DOCUMENT RESUME

ED 131 119	95	TM 005 845

AUTHOR	Cramer, Elliot M.; Appelbaum, Mark I.
TITLE	An Evaluation of Some Methods Used in the National Assessment of Educational Progress. Final Report.
INSTITUTION	North Carolina Univ., Chapel Hill. L.L. Thurstone Psychometric Lab.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	76
GRANT	NEG-00-3-0111
NOTE	131p.
EDRS PRICE	MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS	*Academic Achievement; Analysis of Covariance; *Analysis of Variance; Comparative Analysis; *Groups; *National Surveys; *Statistical Analysis
IDENTIFIERS	*Balancing; National Assessment of Educational Progress; Nonorthogonal Analysis of Variance

ABSTRACT

A recurring problem in educational research has been the adjustment of data to account for initial differences among observed groups of individuals on attributes uncontrollable by the researcher. The procedure called "balancing" is introduced in the National Assessment of Educational Progress report as an adjustment method for this purpose. Since it is apparent that balancing is being used extensively both in the NAEP work and in the analysis of data from state assessments, this research aims at the development of a better understanding of the method and an evaluation of its strengths and weaknesses. The investigation of the nature of balancing required a detailed investigation of the nonorthogonal analysis of variance, the fundamental concepts of marginal means and marginal populations, as well as the investigation of balancing-like data analytic techniques such as "smear and sweep," analysis of covariance, and standardization. It was concluded that the general framework of nonorthogonal analysis of variance encompasses the most useful of the adjustment procedures when used in conjunction with the estimation of weighted marginal means. (RC)

TM

FINAL REPORT

An Evaluation of Some Methods Used in the National Assessment of Educational Progress

National Institute of Education Project No. NEG-00-3-0111

73-2526

Elliot M. Cramer and Mark I. Appelbaum

Psychometric Laboratory

Department of Psychology

University of North Carolina at Chapel Hill

Chapel Hill, North Carolina

1976

The research reported herein was performed pursuant to a grant contract with the National Institute of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or policy.

Table of Contents

SUMMARY

The National Assessment of Educational Progress (NAEP) has had as its purpose the measurement of educational achievement in children and young adults. NAEP Report 7 (1971) is of particular interest and importance in that it characterizes the performance of blacks, of respondents with differing levels of parental education, and respondents from differing types of community. The authors note that the report describes differences as they are and as they would be in particular subgroups if the effects of other characteristics were represented proportionately in each subgroup. Since in a direct comparison between group effects, one characteristic can masquerade effects of another, the method selected for comparing groups is of great importance. For example, on science exercises in the report there is a 20% difference between the extreme affluent suburbs and the extreme inner city. Because of the difference in parental education of the two groups, part of this 20% difference may be "considered to grow out of the difference in parental education." One would wish to compare the two groups as if they were comparable with respect to parental education.

The procedure called "balancing" is introduced in the NAEP report as an adjustment method for this purpose. Little seems to be known about the properties of the method beyond the brief description given in the report. Since it is apparent that "balancing" is being used extensively both in the NAEP work and in the analysis of data from state assessments such as the State Assessment of Educational Progress in North Carolina, the development of a better understanding of the method and an evaluation of its strengths and weaknesses is vital. This has been the principal aim of the research described in this report.

The investigation of the nature of balancing has required a detailed investigation of the nonorthogonal analysis of variance, the fundamental concepts of marginal means and marginal populations, as well as the investigation of balancing-like data analytic techniques such as "smear and sweep," analysis of covariance, and standardization. It has been concluded that the general framework of nonorthogonal analysis of variance encompasses the most useful of the adjustment procedures when used in conjunction with the estimation of weighted marginal means.

The material in this report was prepared by

Mark I. Appelbaum
Elliot M. Cramer
Lyle V. Jones
Scott E. Maxwell
Samuel Peng

Chapter I: Introduction

In surveys one typically describes the ways in which particular groups of individuals differ. One would frequently like to know why the groups differ and whether the differences might be ascribed to other variables which might be modified by educational intervention. The National Assessment of Educational Progress (NAEP), for example, has had as its purpose the measurement of educational achievement in children and young adults. Of particular interest has been the performance of blacks, of respondents with differing levels of parental education, and types of community. NAEP Report 7 (1971) describes differences <u>as they are</u> and <u>as they would be</u> in particular subgroups if the effects of <u>other</u> characteristics were represented proportionately in each subgroup. The method of comparison is of great importance since in a direct comparison of groups the differences in one characteristic may actually be due to another characteristic. The procedure called "balancing" is introduced as an adjustment method for this purpose, apparently for the first time. It is described by the as follows:

"The unadjusted results as reported here and in Report 4 clearly and accurately estimate the differences in achievement between specific groups of children. For example, over all the science exercises, the median percentage difference between 13-year-olds in the Extreme Affluent Suburbs and in the Extreme Inner City is 20% (from Exhibit 6-1). Except for sampling error, this accurately reflects how these two groups differ.

"However, children in the Extreme Affluent Suburb tend, more than children in the Extreme Inner City, to have better educated parents. Because of this lack of balance, part of the difference between these two groups may be considered as growing out of the difference in parental education. Part, also, may be attributable to other factors on which the two groups differ. Some of these factors have been determined for our respondents--their sex, color and the region of residence. Many other possibly relevant factors have not been determined, such as the economic level of the children's parents and the cultural environment in the home.

"It is natural to ask, 'What would the difference between these extreme types of community have been if the distribution of Parental Education, sex, color and region had been the same for both types of community referred to above?' Were it possible to rearrange the world to equate these distributions for each type of community, the effects upon our nation and its schools would be profound. Such rearrangement is not possible. It is usually appropriate to think of the balanced results presented in this report as reflecting the differences we would see in the absence of masquerading by the other four factors. We can be reasonably sure the balanced results do a much better job than the unadjusted results of reflecting such differences."

Apparently the only justification currently available for the use of the method is contained in a ten page appendix of illustrative examples. The basic data treated in the examples are two-way tables of frequency counts giving the number of individuals in a particular cell who have successfully performed on a particular science exercise. This is illustrated in Example 1 where a random sample of 600 individuals is drawn from some well-defined population. The number of cases in each cell is representative of relative number in the population for the particular combination of conditions specified, and the degree of success for that group is estimated by the proportion of respondents giving correct answers to an exercise.

From the two tables, one for numbers of observations and one for numbers of successes, the marginal values are row and column totals.

Example 1

number of observations

		B		
		1	2	
A	1	100	100	200
	2	50	150	200
	3	0	200	200
		150	450	600

number of successes

		B		
		1	2	
B	1	50	30	80
	2	30	60	90
	3	--	100	100
		80	190	270

The problem of concern to the authors of the NAEP Report is that the marginal proportions may not be representative of the underlying effects. There is one sense in which these values are representative; the data are from a well defined population, and the marginal proportions are estimates of proportion of success for that population. However, if one wishes to get at an assumed underlying effect of extreme inner city uncontaminated, say, by the effect of parental education, these marginal values are not representative. Their introduction of balancing was an attempt to obtain representative values.

The NAEP Report notes that interactive differences are not considered and balancing does not adjust for them; and also that, "The deficiencies of balancing are clear; it cannot be the final answer." Balancing will frequently involve estimation in a linear model that is known to be wrong, e.g., when there are interactions present. Also there are other choices of weights, and although other choices do not affect differences between effects, they do affect the absolute magnitudes. We need then to develop a deeper understanding of nonorthogonal ANOVA which will carry over to the interpretation of balanced estimates, as well as providing insight into data analysis more generally. It should be noted that although the National assessment uses medians rather than means for estimation and uses special methods for estimating standard errors, the formulation presented here may perfectly well be used for estimating adjusted effects.

We will show that the method of balancing can be developed in conceptually quite a different way which makes clear that it is a special case of nonorthogonal analysis of variance. The problem of interpreting balanced estimates then can be related to the more general problem of interpreting adjusted effects in the nonorthogonal analysis of variance. This more general problem has been of concern to us since there is not currently a consensus of opinion on the proper methods of analysis for this more general situation. This is reflected by the divergent suggestions we have received from mathematical statisticians regarding the testing of main effects by eliminating both interactions and other main effects, as opposed to eliminating only other main effects. Of course such problems of interpretation arise in regression analysis, too. We have been concerned with this area as well. In a recent article (Cramer, 1972), misuses of regression analysis were discussed, even some that had been published in The American Statistician.

Chapter II: Balancing and the Analysis of Variance

The primary aim of this grant, the explication of balancing in terms of the analysis of variance, is presented in this chapter. It is shown that, without question, balancing is intimately related to classical nonorthogonal ANOVA. The implications for the interpretation of balanced results are presented.

The educational researcher engaged in large scale multifactor survey research may often be faced with a substantial statistical problem whenever the number of observational units is not equal in each and every cell of an experimental or survey design. This situation may arise for a number of reasons ranging from the state of nature to the socio-politics of educational research. For whatever reason the nonorthogonality occurs, the statistical problem remains the same, namely that of being able to estimate the effects of the several states of nature uncontaminated by one another. Simple methods of computing marginal means, marginal percentages, etc. will not yield the desired results.

In an attempt to provide an appropriate method for assessing such effects, Tukey and his associates in the NAEP (1971) studies have offered a method called balancing or the balanced fit. While this method does indeed provide the appropriate estimates of effects under a somewhat restrictive set of assumptions, it is presented in a manner which tends to obscure the meaning of these estimates in relation to well known statistical methods. Indeed, we shall show that the estimation procedure in balancing is nothing more nor less than the estimation procedure in the ordinary Least Square estimation of a nonorthogonal main effects model analysis of variance.

An Example[3] and an Incorrect (but Traditional) Analysis

Let us assume, for illustration, that the following survey had been undertaken - first grade classrooms from three geographical

[3] The data for this example were adapted from NAEP Report 7 (1971).

areas of the country have been randomly selected, 200 from each area, and the method of teaching reading noted for each class - either "phonic method" or "sight method." Each student in each class is given a standardized reading test and the total classroom experience is rated a success if one half or more of the students in the class score at or above their individual age norm on that test. The researcher is interested in assessing the effects of method of instruction and region of residency upon reading skills.

Table 1 shows the number of classrooms (n_{ij}) observed in each cell of the survey. It is apparent from Table 1 that there are three times as many sight reading classes as phonic reading classes and that there are no phonic classes in Region III. Furthermore, the design is unbalanced (nonorthogonal) since the cell frequencies are unequal and there is no constant of proportionality between the numbers in either rows or columns of the design.

Let us assume that the world operates in such a way that the proportions of successes (classrooms in which 50% or more of the students operate at or above their age norm) are as given in Table 2. Within any of the three regions the phonic method is 20 percentage points higher than the sight method, and within either reading method Region I is 10 percentage points higher than Region II which is in turn 10 percentage points higher than Region III.

Now let us suppose that our researcher is "in luck" and the true proportions given in Table 2 exactly reveal themselves in

Table 1

Number of Observations in Cells (n_{ij})

	Phonic	Sight	
Region I		100	200
Region	50	150	200
Region III	0	200	200
	150	450	600

Table 2

True Proportion of Successes in Cells

	Phonic	Sight
Region I	.70	.50
Region II	.60	.40
Region III	.50	.30

Table 3

Observed Number of Successes, k_{ij}, with Observed Proportion, p_{ij}, of Successes in parentheses

	Phonic	Sight
Region I	[illegible] (.70)	50 (.50)
Region II	30 (.60)	60 (.40)
Region III	0 (-)	60 (.30)

Table 4

Marginal Number of Successes, ($k_{i.}$ and $k_{.j}$), Marginal Prop , ($p_{i.}$,$p_{.j}$), and Differential Percentages

	Marginal # Successes	Marginal % Successes	Differential %
Region Marginals	$k_{i.}$	$p_{i.}$	
Region I	120	60%	+15%
Region II	90	45%	0%
Region III	60	30%	-15%
Method Marginals	$k_{.j}$	$p_{.j}$	
Phonic	100	67%	22%
Sight	170	38%	- 7%
Total	$k_{..}$	$p_{..}$	
	270	45%	----

his data, i.e., the number of successes k_{ij} yielding observed proportions p_{ij} in each cell are as presented in Table 3. We can see that in each cell, save Region III phonic classes which are nonexistent, the proportion of successes is identical to that given in Table 2. The researcher is, however, interested in the differential effects of region and method of instruction, and so calculates the number of successes in each row and each column, finding the marginal number of successes given in Table 4. Upon converting to percentages (dividing the marginal number of successes by the marginal totals), he finds the percentages given in the column labeled "Marginal % Successes." He further notes that overall 45% of the classrooms meet the success criteria. Since he is interested in the differential effects of region and instructional method, he then subtracts the total percent success (45%) from each of the marginal percentages yielding the results presented in the column labeled "Differential %." These results indicate that the difference between Region I and Region II is 15% and that between II and III is 15%, while the difference between the instructional methods is about 29%. We know, however, from Table 2 that the effects are actually 10% for each region and 20% for reading method. There appears to be a contradiction.

These results illustrate the problem encountered when the cell frequencies in a design are unequal and disproportional. Clearly we desire an analytic method which accurately reflects the differential effects of the classificatory variables and which can reproduce accurately the observed data from the estimated effects. The naive method illustrated above does neither. The problem is, essentially, that in the disproportional cell frequency case, one effect can masquerade as another. The example given is particularly complex because the estimates of the region effect are confounded with those of instructional method, while simultaneously the instructional method effect is confounded with region effects (i.e., neither set of estimates is free of the influence of the other).

The Balanced Fit and Estimated Effects

The method proposed in NAEP Report 7 (1971) for estimating the effect of one classification variable uncontaminated by the influences of the other in a two way cross-classification has been designated the "balanced fit" by its authors. We find the fundamental principle of the balanced fit stated in the NAEP report as follows:

> We intend to find group effects (expressed in percentages) that, when combined by addition with each other and with the overall percentage of success, give fitted percentages of success that correspond with the actual data in one simple way:
>
> —if we choose any group by a single characteristic, say group A, and if we use the fitted percentages and the actual number of cases to calculate the number of successes for each subgroup that involves (group A), and if we then add these calculated numbers of successes, the total number of successes over all subgroups will be the same as the total actually observed in the data.

Let us take estimated group effects to mean differential row effects, say $\hat{p}_{i.}-\hat{p}_{..}$, and differential column effects, say $\hat{p}_{.j}-\hat{p}_{..}$ where $\hat{p}_{..}$ is the estimated overall proportion of successes. We may then write an expression for the estimated proportion of successes in each cell as

$$\hat{p}_{ij} = \hat{p}_{..} + (\hat{p}_{i.}-\hat{p}_{..}) + (\hat{p}_{.j}-\hat{p}_{..}). \tag{1}$$

Since the observed number of successes for each cell in $n_{ij}p_{ij}$ while the predicted number of successes is $n_{ij}\hat{p}_{ij}$, the basic principle of balancing, that the sum of the observed numbers of successes equal the sum of the predicted number of successes then gives the row conditions

$$\sum_j n_{ij}\hat{p}_{ij} = \sum_j n_{ij}p_{ij} \qquad i=1,2,\ldots,I$$

and similarly the column conditions

$$\sum_i n_{ij}\hat{p}_{ij} = \sum_i n_{ij}p_{ij} \qquad j=1,2,\ldots,J$$

or equivalently

$$\left.\begin{array}{l} \sum_j n_{ij}(p_{ij}-\hat{p}_{ij}) = 0 \\ \\ \sum_i n_{ij}(p_{ij}-\hat{p}_{ij}) = 0 \end{array}\right\} \tag{2}$$

Since there are, in fact, infinitely many solutions to this system of simultaneous equations, two additional conditions are introduced in balancing which make the solution unique

$$\left.\begin{array}{l} \sum n_{i.}(\hat{p}_{i.}-\hat{p}_{..}) = 0 \\ \sum n_{.j}(\hat{p}_{.j}-\hat{p}_{..}) = 0 \end{array}\right\} \tag{3}$$

where $n_{i.}$ and $n_{.j}$ are the numbers of observations in the ith row and jth column, respectively. Applying the constraints in (3) to (1), we see that

$$\hat{p}_{i.} = \frac{\sum_{j} n_{.j}\hat{p}_{ij}}{n_{..}} \quad \text{and} \quad \hat{p}_{.j} = \frac{\sum_{i} n_{i.}\hat{p}_{ij}}{n_{..}}$$

and that the $p_{i.}$ and $p_{.j}$ are marginal proportions using the weights $n_{.j}$ for rows and $n_{i.}$ for columns. As the NAEP Report notes, these conditions, (2) and (3), are sufficient to uniquely define the group effects in (1) and hence to uniquely define the estimated marginal proportions $\hat{p}_{i.}$ and $\hat{p}_{.j}$. We shall now show that (1), (2), and (3) have exact parallels in the nonorthogonal analysis of variance.

Estimation in the Nonorthogonal Analysis of Variance

Readers familiar with the analysis of variance (ANOVA) will recognize certain similarities between the survey design of the example and designs often analyzed by ANOVA techniques. We shall now show that the estimates of differential effects obtained from the balancing procedure are exactly the same as those produced by a nonorthogonal analysis of variance of a main effects model. It should be emphasized that we are dealing, at this time, with estimation in the ANOVA model, not the tests of significance which are more commonly seen in ANOVA applications.

This important relationship between balancing and ANOVA will be more easily seen if we adapt our notation and terminology to that commonly employed in the ANOVA context. In this case we are dealing with a two-way cross-classification, often called a two-way factorial, with unequal and disproportional cell frequencies (a nonorthogonal factorial design). We now consider our first classification (factor), labeled A, to have I levels and the second classification (factor), labeled B, to have J levels. We will use the symbol y_{ijk} to represent the score of the kth classroom under the ith level of A and the jth level of B. Let y_{ijk} equal one for a success (if 50% or more of the students in the class score at or above their age norm) or zero for a failure; y_{ijk} is then a binary random variable. The cell mean

$$\bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$$

is simply the observed proportion of successes in the i,jth cell and will correspond to p_{ij} in our earlier notation.

In the estimation phase of the analysis of variance employing a main effects model, one attempts to predict the i,jth cell mean through the linear model

$$\hat{y}_{ij} = \hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j$$

where the $\hat{\alpha}_i$ may be thought of as the estimated differential effect of the ith level of A, $\hat{\beta}_j$ as the estimated differential effect of the jth level of B, and $\hat{\mu}$ as a general or average effect about which the differential effects operate. In the analysis of variance we estimate the values of these parameters according to the Method of Least Squares, i.e., so that the sum of squared deviations of the observed scores from the predicted scores is a minimum. If we let $\hat{p}_{ij}$ indicate that value of the cell mean predicted from the Least Squares estimates of the parameters for the i,jth cell, writing

$$\hat{p}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j, \qquad (4)$$

we may obtain the least squares estimates of the unknown parameters by minimizing

$$S = \sum_i \sum_j n_{ij} (p_{ij}-\hat{p}_{ij})^2$$

Since we are predicting cell means, we weight each cell by the number of observations in that cell. To minimize S we may differentiate with respect to μ, α_i and β_j to obtain

$$\left.\begin{aligned} \sum_i \sum_j n_{ij}(p_{ij}-\hat{p}_{ij}) &= 0 \\ \sum_j n_{ij}(p_{ij}-\hat{p}_{ij}) &= 0 \\ \sum_i n_{ij}(p_{ij}-\hat{p}_{ij}) &= 0. \end{aligned}\right\} \quad (5)$$

The equations in (5), usually referred to as "normal equations" in the ANOVA context, do not themselves uniquely define $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$. They do, however, yield unique values of $\hat{p}_{ij}$; that is, any set of values which is a solution of (5) will yield the same values of the $\hat{p}_{ij}$. Substituting into (4), it follows that $\hat{\alpha}_i - \hat{\alpha}_k = \hat{p}_{ij}-\hat{p}_{kj}$ is uniquely defined regardless of which particular set of $\hat{\alpha}$'s are used. This result is easily generalizable to the fact that contrasts in the unknown parameters are unique for any solution of (5).

In order to obtain computationally unique solutions for these parameters it is the usual practice in the analysis of variance to further constrain the system by the condition that

$$\sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = 0 .$$

While this is the most commonly employed set of constraints, any other set of constraints will work equally well and will not change the meaning of the resulting solution so long as one considers only contrasts in the parameters. Given this freedom of choice, we prefer to use the constraints

$$\Sigma n_{i.}\hat{\alpha}_i = \Sigma n_{.j}\hat{\beta}_j = 0 \quad (6)$$

which are those commonly employed in the nonorthogonal ANOVA (see for example Winer, 1971).

We now note [illegible] equation set [illegible] is exactly the same as the balancing equation set (2). If we further equate

$$\hat{\mu} = \hat{p}_{..}$$
$$\hat{\alpha}_i = \hat{p}_{i.} - \hat{p}_{..}$$
$$\hat{\beta}_j = \hat{p}_{.j} - \hat{p}_{..}$$

equation (6) is exactly the same as the balancing equation (3). Our basic ANOVA model (4) may then be written as

$$\hat{p}_{ij} = \hat{p}_{..} + (\hat{p}_{i.} - \hat{p}_{..}) + (\hat{p}_{.j} - \hat{p}_{..}) \qquad (7)$$

so that we have an exact equivalence between (7) and (1) and hence between balancing and nonorthogonal ANOVA.

Substituting (6) and (4) in (5) we can also show that

$$\hat{\mu} = \frac{\Sigma\Sigma n_{ij} p_{ij}}{\Sigma\Sigma n_{ij}} = \hat{p}_{..}$$

as is assumed in the NAEP Report. Thus, we see that the balancing equations are but a special case of the least squares equations of a nonorthogonal ANOVA in a main effects model, and, in this sense, the two are equivalent.

The correspondence between the balancing algorithm and that of the nonorthogonal analysis of variance makes possible the use of standard ANOVA programs which properly analyze nonorthogonal designs (e.g., Cramer, 1967) for obtaining balanced fits. Since current usage of the balancing technique has been limited to obtaining estimated cell means and contrasts in main effect parameters there is no particular concern with the constraining system employed since these solutions are invariant with respect to the constraining system. If, however, one wishes to obtain the estimates of the parameters themselves it would be necessary to employ an ANOVA program which allows for the specification of the constraints given in (6). Cramer's (1967) program, for one, allows such a specification.

Generalization to Higher Order Classifications and to Data Type Other Than Proportions

It can be shown that the generalization of balancing to more than two classifications is equivalent to estimation in a higher order nonorthogonal ANOVA with a main effects additive model. Thus, it is possible to produce estimates of effects balanced for more than one interfering variable. Furthermore, there is no need to restrict estimates to those of proportions. Since balancing does not uniquely require data in the form of proportions (although it is nearly always so illustrated), one could equally well use the cell means of continuous response data in order to obtain balanced estimates of differential effects.

The Interpretation and Meaning of Balanced Estimates

When dealing with the nonorthogonal analysis of variance (of which balancing is just a special case) careful attention must be given to the meaning and interpretation of estimates and tests of significance. Appelbaum and Cramer (1974) have discussed the problems involved in tests of significance at some length. The critical problem in the nonorthogonal case is that the effects of the several states of nature upon the dependent variable in general cannot be estimated or tested separately;

they are inherently confounded. The exact manner in which such data are treated has very profound effects upon the meaning of the resulting estimates and the ways in which they may be interpreted.

A thorough understanding of the nature of the estimation procedure used in the nonorthogonal analysis of variance (and hence balancing) may be best facilitated from a consideration of marginal means rather than of the estimates themselves. The parameters estimated in ANOVA (the effects) are defined as functions of certain population means. It is clear that all the information available for the estimates of effects is included in the estimates of the marginal means. Recalling that the p_{ij} are themselves cell means, the differences between effects which are of particular interest can also be expressed as differences in marginal means. For instance, if we were interested in the differences between effects of the first two levels of the A classification, we would be interested in $(\hat{\alpha}_1-\hat{\alpha}_2)=(\hat{\mu}_{1.}-\hat{\mu}_{..})-(\hat{\mu}_{2.}-\hat{\mu}_{..})=(\hat{\mu}_{1.}-\hat{\mu}_{2.})$.

In the process of selecting the way in which we produce the estimates of these differences, we are actually making two quite different (and to some extent independent) decisions. One is fundamentally a question of what it is that we wish to estimate; the second is a decision of how to estimate that which we have decided to estimate. The first is a question of weighting; the second is a question of models and adjustments.

When one does an experimental study, be it a true experimental manipulation or a survey, one considers that each cell of the design represents a random sample from some conceptual population. In a two way classification, the true population mean of one such conceptual population would be represented as μ_{ij}. It is these and only these basic populations which have an invariant meaning defined by the basic design of the experiment. When we begin to introduce the concept of a marginal mean (as we must when we talk of effects) we are adding a new conceptual dimension, for marginal means are weighted linear combinations of the basic population means. The way we choose to weight the population means in effect defines the marginal populations from which the estimates will be obtained. It must be understood that (1) marginal populations have no reality beyond the nature of the basic populations and the way in which they are combined, and (2) that the meaning of the estimated effects will depend upon what weights are selected (i.e., the weights will determine what is being estimated).

A weighted mean is any linear combination of observations with positive coefficients which sum to one. There are, of course, many different sets of coefficients with this property, implying that there are many different conceptual marginal populations which could be defined. There are, however, three basic types of weightings which might be employed for a two way design: (1) equal weights, (2) singly subscripted weights, and (3) doubly subscripted weights. In order to understand the nature of these three weighting schemes, consider for the moment the situation in which we know the estimated population means $\hat{\mu}_{ij}$ for each and every cell in a two way design. Consider the construction of row marginal means with each of the three weighting schemes.

In the first case, all weights employed would be equal so that the marginal population mean for the ith row of the experiment would be

$$\hat{\mu}_{i.} = \sum_j \hat{\mu}_{ij}/J.$$

This type of marginal mean is usually referred to as an unweighted marginal mean. In this case, each of the basic populations is treated as being identical to all other populations in its contribution to the marginal populations. In the second case the weights carry only a single subscript yielding row means of the form

$$\hat{\mu}_{i.} = \sum_j w_j \hat{\mu}_{ij}.$$

The several basic populations entering into the row marginal mean are differentially weighted, but the weights are the same for every row. These marginal means will, in general, be different from the unweighted means. For the third case the weights for each row will sum to one but they will differ from row to row. In this case the marginal mean for the ith row will be

$$\hat{\mu}_{i.} = \sum_j w_{ij} \hat{\mu}_{ij}.$$

A question which must concern us is "for what situation will we want to use which of the various weighting schemes?" If in the example considered we were interested in estimating the differences between the two reading methods <u>as they are used throughout the country</u>, we would be interested in differential effects based upon weighted marginal means, where the weights reflect the number of classes using a particular reading method in a particular region of the coutry (a doubly subscripted weighting system). If, on the other hand, we were interested in estimating the differences between the two reading methods as if they were equally used throughout the country, we would be interested in differential effects based upon equal weights. A third possibility would be to assume that the use of both methods was in proportion to the population in the various regions. This would imply that the same weights would be used for both methods, i.e., singly subscripted weights. The choice of the weighting system is entirely up to the investigator, but the choice is not a trivial one. The selection of the weighting system basically defines what it is that the researcher is referring to. One further refinement on the nature of the weights in the case of balancing will be added shortly.

The Nature of the Weights

Up to this point, nothing has been said about the nature of the weights themselves. In practice the weights may represent any conceptual entity which the researcher deems important, say the relative cost of a treatment, the current social importance of a particular segment of the population, etc. Surely the most common weights by far are the relative sample sizes. Insofar as the observed cell frequencies represent (are proportional to) the actual population sizes, weighting by the cell frequencies may be logically sound. In those cases where the observed cell frequencies do not reflect any true state of nature, or when the populations are considered to be infinite, such a weighting scheme can make little if any sense.

At this point a word of caution seems to be in order. It is important to distinquish between weights as we have defined them above and the coefficients of the "normal equations" in (5) and in the constraints of (6). The weights are defined for the purpose of constructing marginal populations; the coefficients of (5) and (6) are the result of the Least Squares criteria and are completely independent of considerations relating to the definition of marginal populations.

The Problem of Estimation

Having decided upon a weighting scheme and thereby defining marginal populations and a potential set of effects to be estimated, one is left with a second, although not totally independent question of how to do the estimation. Clearly, if we possess unbiased estimates of the individual population means we can easily obtain unbiased estimates of the marginal means no matter how they are defined. Since linear combinations of unbiased estimates of population means produce unbiased estimates of the same linear combination of population values, we may always obtain the desired unbiased estimates. Thus, the problem of estimation reduces to the problem of how to produce unbiased estimates of the individual population means.

Whenever one establishes estimates of parameters, say population means, one is always operating within the context of a model; the nature of the obtained estimate depending upon the model in which it is estimated. In the two way classification scheme there are five reasonable[4], but different models which might be used to estimate a typical population mean μ_{ij}. These are:

i) $\mu_{ij} = \mu+\alpha_i+\beta_j+\alpha\beta_{ij}$ (the interaction or cell means model)

ii) $\mu_{ij} = \mu+\alpha_i+\beta_j$ (the two main effects model)

iii) $\mu_{ij} = \mu+\alpha_i$ (the main effects A model)

iv) $\mu_{ij} = \mu+\beta_j$ (the main effects B model)

v) $\mu_{ij} = \mu$ (the grand mean model).

When an experimental design is nonorthogonal and when the Least Squares estimation procedure is used for obtaining the estimates, very different estimates of the population means will obtain for estimation in the different models and, as a consequence, different estimates of the marginal means and differential effects will result.

If we choose to estimate the individual population means in the first model (the interactive or cell means model), the ordinary cell mean, $\bar{y}_{ij}$, is obtained as the estimate. $\bar{y}_{ij}$ is always an unbiased estimator of the population mean μ_{ij} without regard to which model obtains in nature. If, however, one of the simpler models should be the true model, the variance of the $\bar{y}_{ij}$'s will be larger than the variance of the unbiased estimator resulting

[4] Some authors have suggested other possible models, e.g.

$$\mu_{ij} = \mu+\alpha_i+\alpha\beta_{ij}.$$

Problems involved with such models have been discussed elsewhere (e.g. Appelbaum & Cramer, 1974) and are not considered here.

from estimation in the correct model. Thus, $\bar{y}_{ij}$, while always providing an unbiased estimator, will not be the minimum variance unbiased estimator unless there is truly an interaction between the classification factors.

When estimation proceeds from the second model (the two main effects model), one obtains the balanced fit estimates of the population means (or equivalently the main effects ANOVA estimates). These estimates are unbiased only when one of the non-interaction models (ii, iii, iv, or v) holds in nature and will be minimum variance unbiased only when model ii holds. Thus, estimates from model ii, often called adjusted estimates, are appropriate only in the non-interactive case.

In a similar fashion, estimates based on models iii, iv, and v will be unbiased estimates only when the corresponding models obtain. These estimates provide minimum variance unbiased estimators only when the particular model holds.

One is free to select an estimation scheme based upon one's belief in the state of nature, but one must always remember that this choice will simultaneously affect the resulting estimates both in terms of their unbiasedness and variance. In order to obtain unbiased minimum variance estimates one must estimate in the model corresponding to the true state of nature. Should the model selected be too simple relative to the true state of nature, the estimates will, in general, be biased; should the model be too complex, the estimates will not be minimum variance.

The Intersection of Weighting Schemes and Estimation

Any procedure for obtaining estimates of effects in an n-way layout can now be viewed as the intersection of a weighting scheme and an estimation procedure, and its properties may be better understood by examining the consequences of the individual components. Balancing is, in this view, the intersection of a singly subscripted weighting system with model ii estimation (two main effects model). Thus, the balanced estimates of differential effects are estimates obtained employing singly subscripted weights for each population and by assuming no interaction among the classification dimensions.

It is, however, possible to view balancing as the intersection of equal weights and model ii estimates. This indeterminacy occurs because of an interaction between the weighting system and estimation system employed. This result, which has major implications for the interpretation of the balanced fit may be understood more easily by returning to our initial example.

The NAEP investigators discuss balancing in terms of making comparisons between two groups as if the groups were identical to one another in terms of their compositions on other (interfering) variables. This goal clearly implies a singly subscripted weighting system. In terms of our initial example, this amounts to asking about the differences in reading method as if they were used in the same proportion in all three regions of the country. Thus we would be interested in the column marginal difference with the rows weighted the same for both columns; i.e. we are interested in

$\hat{\mu}_{.1} - \hat{\mu}_{.2}$ where $\hat{\mu}_{.1} = v_1\hat{\mu}_{11} + v_2\hat{\mu}_{21} + v_3\hat{\mu}_{31}$

and $\hat{\mu}_{.2} = v_1\hat{\mu}_{12} + v_2\hat{\mu}_{22} + v_3\hat{\mu}_{32}$

The v_i's are the weights to be applied to each region and may, for instance, reflect the relative sizes of the region in terms of the number of first grade classrooms. We may now write

$$\hat{\mu}_{.1} - \hat{\mu}_{.2} = \widehat{\mu_{.1} - \mu_{.2}} = (v_1\hat{\mu}_{11} + v_2\hat{\mu}_{21} + v_3\hat{\mu}_{31}) - (v_1\hat{\mu}_{12} + v_2\hat{\mu}_{22} + v_3\hat{\mu}_{32}) = v_1(\hat{\mu}_{11} - \hat{\mu}_{12}) + v_2(\hat{\mu}_{21} - \hat{\mu}_{22}) + v_3(\hat{\mu}_{31} - \hat{\mu}_{32}).$$

The implication of the noninteractive model employed by the balancing system, however, is that all row differences must be equal across columns and that all column differences must be equal across rows. Therefore, $(\hat{\mu}_{11} - \hat{\mu}_{12}) = (\hat{\mu}_{21} - \hat{\mu}_{22}) = (\hat{\mu}_{31} - \hat{\mu}_{32}) = \Delta$. Thus $\widehat{\mu_{.1} - \mu_{.2}} = v_1\Delta + v_2\Delta + v_3\Delta = (v_1 + v_2 + v_3)\Delta$. We further note that the weights must be chosen to sum to 1 by their very definition and hence $\widehat{\mu_{.1} - \mu_{.2}}$ must equal Δ. The implication of this result is that it makes absolutely no difference how the subpopulations are weighted in constructing the marginal population in the balanced solution as long as they are weighted the same for each population. This implies that the true relative sizes of, say, regions of the country do not enter into the assessment of the methods difference. Since equal weights are but a special case of singly subscripted weights they could equally well be used. We therefore conclude that in balancing it makes not the least bit of difference whether we equally weigh the subgroups or weight them differentially in the sense of singly subscripted weights.

The "traditional but incorrect analysis" presented as the first example in this paper is an example of an unweighted scheme with model iii estimates obtained for the rows and model iv estimates obtained for the columns. The incorrectness of this analysis arises from the fact that we are applying a one main effect model when indeed there are two main effects. The row marginal means are obtained from a model $\hat{p}_{ij} = \hat{\mu}+\hat{\alpha}_i$ where $\mu+\hat{\alpha}_i=\sum_j n_{ij}p_{ij}$ while the column marginal means are obtained from a model $\hat{p}_{ij}=\hat{\mu}+\hat{\beta}_j$ where $\mu+\hat{\beta}_j=\sum_i n_{ij}p_{ij}$.

Conclusion

It has been shown that the estimation procedure employed in balancing is nothing more or less than the Least Squares estimation of effects in a nonorthogonal main effects model analysis of variance. In assessing the appropriateness of this method of analysis for a particular study, one must consider the appropriateness of the two component elements; first that of the weighting scheme employed and second that of the use of the main effects model.

Balancing, it has been seen, can be viewed as employing either singly subscripted weights or equal weights; the results being invariant to this selection. It should be noted that these are not the only possible schemes, nor the ones necessarily desired. One could alternatively use the cell estimates obtained for the solution of the balancing equations, but then use unequal weights to define marginal population means. The selection is up the researcher and depends only upon what it is that he wishes to estimate.

The assessment of the use of the main effects model is a somewhat more difficult issue due to the fact that the appropriateness of the model depends upon what is true in nature, not simply upon what we would like to be true. In many ANOVA applications this is not a particular problem for one often tests the significance of the interaction prior to estimation in order to determine what is the proper model. In using balancing, however, one is at the outset <u>assuming</u> that the main effects model is appropriate. The basic implication of this assumption is that we are assuming no differential effect of one state of nature conditional upon another. In our example we are, for instance, assuming that the difference between the efficiency of phonic and sight methods of reading instruction is the same in each and every region under study. The tenability of the non-interaction assumption is, of course, completely dependent upon the particular study under consideration and no general rules can be formed for saying *a priori* when the assumption holds. There may be certain circumstances under which the additive model is appropriate, but it would seem, in general, to be a dangerous assumption to routinely employ.

Chapter III: The Nonorthogonal ANOVA

Central to the understanding of balancing is the concept of the nonorthogonal ANOVA. The following chapter serves to illuminate the fundamental concepts of the nonorthogonal design and to resolve a number of the controversies surrounding this general topic.

The nonorthogonal, multifactor analysis of variance (ANOVA) is, perhaps, the singly most misunderstood analytic technique available to the behavioral scientist save factor analysis. Standard textbooks all but ignore it, or, when they do consider this case, bury it in such confused mathematics or approximations as to make it barely understandable to even rather statistically sophisticated researchers. Recent journal articles (e.g., Joe, 1971; Overall and Spiegel, 1969; Rawlings, 1972; Werts and Linn, 1971; and Williams, 1972) have attempted to clarify the situation and set guidelines for the analysis of nonorthogonal, multifactor experiments, but have, in our opinion, done neither. These papers have, again, confused the issues with unnecessary mathematical proofs, with antiquated "approximate" methods, and with the implication that somehow nonorthogonal designs are special cases to be avoided at all costs. So strong is the belief that there is something inherently "difficult" or "strange" about the nonorthogonal case that experimenters will, on occasion, go to unusual lengths, such as randomly discarding data from selected cells, in order to achieve an orthogonal design.

We wish to argue that there is no *conceptual* difference between orthogonal and non-orthogonal ANOVA and that, indeed, the orthogonal design is a special, and occasionally artificial, case of the more general non-orthogonal design. By approaching the entire issue of the analysis of variance as one of model comparisons the special problems encountered in the non-orthogonal case are rather easily understood and resolved. The closely related problem of deletion of variables in multiple regression analysis has been discussed by one of the authors (Cramer, 1972). By treating the problem as one of comparisons of linear models he has resolved the issue in a clear and fairly obvious manner. We believe that a similar approach with non-orthogonal designs will lead to the same resolution

The easy access to sophisticated computer programs which perform the analysis of variance by a general linear model approach (e.g., MANOVA; Cramer, 1967) makes the computations for this method of dealing with non-orthogonal, multifactor designs possible and eliminates, in most cases, the need and desirability for "approximate" solutions.

Terminology and basic concepts

Before proceeding with a detailed discussion of non-orthogonal analysis of variance, it is necessary to clarify some of the terminology and concepts that are fundamental to these analytic techniques. A *non-orthogonal design* refers to any experimental design in which the numbers of observations are not equal in each and every cell. This definition encompasses even designs that are traditionally classified as proportional and includes designs that are not complete factorial. Insofar as an experimental design may be considered a partially complete factorial design (e.g., Randomized Blocks, Latin Squares, nested or hierarchical designs, etc.), the principles discussed in this paper apply.

We shall use the term method to refer to the estimation procedure which we shall assume to be the Method of Least Squares. The concepts developed and discussed in this paper apply only to Least Squares Analyses and should not be applied to non-exact approaches such as the Unweighted Means Analysis (Winer, 1971; 445-449) nor to cases which employ some other method of estimation of effects.

By an experimental design we shall mean the plan of the experiment determined by the experimenter on the basis of his conception of some idealized state of nature. The minimum requirements for an experimental design are the specification of the experimental factors to be manipulated and the plan for random assignment of experimental units to treatments, including both the sampling plan and the determination of the number of units per treatment. It is the experimental design which implicitly specifies a set of possible models or idealized states whose appropriateness we shall attempt to assess.

Hypotheses or tests of hypotheses are, in essence, comparisons of various models. It is fundamental that one understand that, within the analysis of variance, one is always trying to assess the appropriateness of one model in comparison to another one. To stress this point, we shall often refer to significance tests as model comparisons. Unfortunately, the standard approaches to the analysis of variance in most introductory courses overlook this consideration and have led to much unnecessary confusion.

Finally, one must carefully consider those situations which may produce a non-orthogonal design. First there is the case where the design is intentionally planned as non-orthogonal and is executed as planned. Such designs are reasonable and may be preferred in cases where contrasts of particular cells are desired, or where greater precision of estimation is required in some cells than in others. Similarly, some experiments, particularly those involving concommitant variables as factors in the design, may be planned as non-orthogonal in order to allow naturally occurring differences in cell frequencies to manifest their effects in the resulting tests. While these designs are rarely encountered in psychological studies, they do have applications and present no particular difficulty in terms of "proper" analysis and interpretation. The discussion of the non-orthogonal analysis of variance which follows is directly applicable to those designs. The second, and far more common, case occurs when a design (orthogonal or non-orthogonal) is not executed as planned. That is, once the random assignment of experimental units to treatments has been made, data are not obtained on some units. In this second case, one of two different situations may have occurred and, depending upon which is true, rather different approaches are required. It may be that the "cause" of the loss of experimental units is a random phenomenon or one unrelated to the experimental treatments. Death of experimental animals, "no-show" of college subjects, etc. often may be viewed as essentially random phenomenon. Again we have no particular problem for we are, in effect, left with a random sample of a random sample which is itself a random sample, and the methods of non-orthogonal analysis of variance to be discussed still apply.

The situation that may cause considerable difficulty is when the "cause" of loss of experimental units cannot be considered a random phenomenon (e.g., it may be related to the experimental treatment). This situation may be obvious, as when the combination of treatments cause the death of some experimental units; or it may be subtle, as when one set of treatment combinations are run late in the afternoon causing an increase in the no-show rate. In such a case, there would seem to be no remedy short of pretending that the missing observations are random, and hoping that the results will be reasonable. Perhaps the definitive statement was made by Cochran and Cox (1957, p. 82) when they observed that the only complete solution to the problem of missing data is not to have any. The following method leads to correct analyses and interpretation of designs which are (1) planned as non-orthogonal or (2) which become non-orthogonal due to the random processes of nature.

Models and the method of least squares.

Having decided, either by choice or default, to employ the method of least squares and having determined the design of the experiment on the basis of a belief as to the nature of the world (in some idealized sense), one is left only with the selection of possible models and model comparisons. We first note that the models selected are logically independent of the observed numbers of observations per cell. While obviously the analysis will be affected by the cell frequencies, the experimenter is free in the design of the experiment to choose the numbers of observations per cell, constrained only by considerations of efficiency and convenience. The models themselves, the representations of our belief in the nature of the world, are not expressed in terms of the number of units in any subpopulation--indeed, the models being considered are usually in terms of infinite subpopulations. Since the model itself is free of population size, the cell frequencies can hardly matter in terms of the correctness of the model. But then why all of the concern about non-orthogonal analysis?

The problem of non-orthogonal analysis really occurs at the level of model comparisons and proper interpretation of the results of such comparisons. As we shall see, the difficulty arises from the methods available to assess the "correctness" of the several models being compared.

One-way ANOVA

Let us first consider a k-group one-way ANOVA with n_j observations in each group; a design which is usually thought to offer no problems, even with unequal cell frequencies. It is our goal to make inferences concerning the population means in the several treatment populations. These inferences will be based upon the observed sample means, the best unbiased estimates of the population means if we make no additional assumptions about the populations beyond those of normality and homogenity of error variance.

One-way ANOVA is commonly treated as the comparison of two models

$$(I)\quad Y_{ij} = \mu+\alpha_j+e_{ij}$$

$$(II)\quad Y_{ij} = \mu+e_{ij}$$

If model II is the correct model, the means for the several populations must be exactly the same and the best unbiased estimate (the least squares estimate) of each of them is the common mean of all the observations. This estimate has variance $\sigma^2/\Sigma n_j$, where σ^2 is estimated by the common within-cell variance.

If model I is correct, the best unbiased estimate of any population mean is the sample mean for that population which has variance σ^2/n_j. The best unbiased estimate of any difference (contrast) in the population means is the difference (contrast) in the sample means. This is true regardless of the number of observations obtained from any population since knowing one population mean tells one nothing about any other population mean. (Note that if Model II is correct, estimates of means obtained from Model I are unbiased but have variances which are larger in the ratio $\sum_{k=1} n_k/n_j$).

The number of observations obviously does effect the variance of the estimates of population means and must also affect the power of any tests of significance. For any two groups (j and k) the variance of the mean differences is the weighted sum of the variances (of means) $\sigma^2/n_j+\sigma^2/n_k$. If the total number of observations for the two groups is held constant, this variance is a minimum when the cell frequencies are equal and the power of the test of the difference of these population means will be a maximum in this case. Similarly it can be shown that the power of the test of equality of all the population means is a maximum when all the cell frequencies are equal. Thus the effect of non-orthogonality in the one-way ANOVA is in terms of the power of the test--not in the obtained estimates nor in the test of their significance.

Two-Way Analysis of Variance

The situation is not nearly so simple when we move to the case of factorial experiments. The additional problems encountered in the factorial case are illustrated by the following example, intentionally constructed to represent an extreme case.

Consider a two-way ANOVA for which we have observed the cell means $\bar{x}_{ij}$ with the cell frequencies n_{ij} as given in Table 1. Assume that the estimated within-cell (error) standard deviation is 15 in each cell (i.e., $MS_{error}=225$).

Table 1

Cell Means and Frequencies for Two-way Example

Cell Means, $\bar{x}_{ij}$

		B	
		1	2
A	1	10	10
	2	20	20

Cell Frequencies, n_{ij}

		B	
		1	2
A	1	25	2
	2	2	25

As an exercise, let the reader consider the answers to the following questions before proceeding further: (1) What can one say, given the above information, about the presence of any main effects or interactions in this experiment? (2) Given the answer to this question, further consider what one would suppose to be true of the populations?

In our experience, relatively sophisticated psychologists and graduate students will not necessarily answer these questions in a consistent manner. We believe that the customary training in psychological statistics will lead many to base their answers to the first question on the means alone judging that there is an A effect but no B effect or interaction. The obvious inequality in the numbers of observations per cell will be troubling and the sophisticated respondent will certainly observe, in response to the second question, that the main diagonal means are much more stable than the off-diagonal means.

If one asks the further question, "What would and should an ANOVA analysis tell you about the true populations?", we are at the heart of the problem of non-orthogonal ANOVA. Surely any ANOVA, orthogonal or not, must give information about the population means. It is not reasonable that unequal numbers of observations in cells will alter the character of this information, although it will certainly alter the precision of any statements.

Looking at the sample means of Table 1 it is apparent that if the population means are the same as these sample means (and this is our best guess), there is only an A effect present. Our statements must, however, take into account the sampling variability of these sample means. Consider, for a moment, the 95% confidence intervals (Table 2) which might be generated about the four observed sample means. Since the samples themselves are independent random samples from four possibly different populations, the confidence intervals are, in the same sense, independent. From these confidence intervals

Table 2

95% Confidence Intervals on Sample Means[1]

		B	
		1	2
A	1	$3.97 \le \mu_{11} \le 16.03$	$-11.32 \le \mu_{12} \le 31.32$
	2	$-1.32 \le \mu_{21} \le 41.32$	$13.97 \le \mu_{22} \le 26.03$

[1] These confidence intervals are based upon the pooled MS error with 50 d.f.

we may see it is reasonable that our sample could have come from a set population with any of the following patterns of means (Table 3). These are but

Table 3

"Reasonable" Population Means

B	
1	2
10	10
20	20

		B	
		1	2
A	1	10	20
	2	10	20

		B	
		1	2
A	1	10	20
	2	30	20

three among many possible sets, but notice that we would consider the first as one in which there was a main effect of A, but no B or AB effects; the second would be considered as an example of a main effect for B but no A or AB effect; while the third would be indicative of a situation with interaction and a main effect.

We thus believe that the conclusion one should logically draw from these sample values is that there are some effects, but that the data do not permit a definitive statement as to which. We further believe that a proper ANOVA should lead one to draw such conclusions.

An incorrect "approximate" analysis

Let us now consider an erroneous analysis which we believe many psychologists might be inclined to perform. This is an analysis of each factor collapsing over the other. Although it does have some intuitive appeal and indeed may be useful in conjunction with other analyses, it will, in general, lead to incorrect conclusions about the population means when used alone.

Suppose we collapse the design given the Table 1 over the B classification leaving us with two levels of A with mean values of 10 and 20 as shown in the marginal values of Table 4. A one-way analysis would then lead to the conclusion that there is a significant main effect of A (p=.017). If we then collapse over levels of A we have the levels of B with means of 10.7 and

Table 4

Means collapsed over Classification

		B		
		1	2	
A	1	10	10	10
	2	20	20	20
		10,7	19,3	

19.3 suggesting a B effect (p=.042) as well as an A effect. We would call these analyses, respectively "A ignoring B" and "B ignoring A". The use of the phrase "A ignoring B" is meant to indicate that in our two-way table we "ignore" the B classification and treat the design as if it were only a one-way classification with levels of A. Observations for a given level of A are considered replicates regardless of whether or not they correspond to the same level of B (that is, we assume no B or AB effects). When there is no B or AB effect, the observations at the several levels of the collapsed factor are, in effect, replicates since the variability between levels of B is of the same order of magnitude as the variability within a level of B. If however, in a non-orthogonal design, there is a B or an AB effect, the estimated magnitude of the A effect (ignoring B) will, in general, be affected by the number of observations in the cells and does not represent an unbiased estimate of any population value. Only when there are equal numbers of observations in the cells will the estimate of the magnitude of the A effect be unaffected by the number of observations in the presence of a B or AB effect.

In general, when we are estimating the magnitude of effects, we may safely ignore other effects in the design only when they are null or when their estimates are independent of the effects in which we are interested. The first condition (that of null effects) depends only upon the state of nature; the second (independence of estimates) depends only upon the actual design of the experiment. The conclusions drawn from this "ignoring" analysis of Table 1 will be incorrect under the (plausible) assumption that there is only one main effect in the population responsible for the results.

For the general non-orthogonal case a different method is necessary in order to estimate treatment effects without bias and to provide unbiased tests of significance. These are tests of "A eliminating B" and "B eliminating A" with corresponding estimates of the effects. In essence these are tests which take into account and eliminate the confounding effects of other factors when they are present. Thus a test of "A eliminating B" "removes" any confounding effects of factor B. If there is no B effect (i.e., it is null in the population) or if the design is orthogonal, there is no confounding due to B and nothing to eliminate; hence the test will be identical to that of "A ignoring B". The test of "A eliminating B" answers the question: given the possibility of a B effect is there evidence for an A effect *in addition* to any B effect which might be present. On the other hand, the test of "A ignoring B" in general answers the question: Is there any evidence for an A effect assuming there is no B effect or ignoring it if it is present. The estimate of the A effect corresponding to the test of "A eliminating B" is unbiased regardless of the existence of any B effect or of orthogonality in the design. It is always the "correct" estimate.

Model Comparisons and Tests of Effects

The more general "eliminating" method described above involves fitting a model allowing for both A and B effects and then comparing the fit (i.e., the quality of the model) to the fit of a model omitting one or more of the effects. For example, consider the following models which "predict" the response of a subject in the ij cell of a two factor design

$$\begin{aligned}
&\text{I.} && Y_{ij}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon\\
&\text{II.} && Y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon\\
&\text{III.} && Y_{ij}=\mu+\beta_j+\varepsilon\\
&\text{IV.} && Y_{ij}=\mu+\alpha_i+\varepsilon\\
&\text{V.} && Y_{ij}=\mu+\varepsilon
\end{aligned}$$

Model I is the most complete model for a two factor design; it allows for an overall level (μ), an effect dependent upon the level of factor A (α_i), an effect dependent upon the level of factor B (β_j), and an interactive effect (γ_{ij}) dependent upon the joint, non-additive effect of the combination of the ith level of factor A with the jth level of factor B. The other models are obtained from the first by dropping the interaction term and possibly one or both of the main effects. Those accustomed to only orthogonal ANOVA will be inclined to regard model I as capable of providing the parametric estimates needed for the other models, but this is not so in general. Each model represents a separate least squares estimation problem and may provide different estimates of the parameters involved. Only in the case of orthogonality will the estimated parameters for the different models be necessarily the same. Likewise, it is only for the orthogonal case that the estimated parameters within a model will be statistically independent (unconfounded) of one another. This is the real meaning of orthogonality.

We would begin the analysis of a two-way factorial, either orthogonal or non-orthogonal, with the test of interaction. Our feelings of parsimony dictate a preference for a main effects model if it is consistent with the data and so we would wish to compare a model allowing for main effects and interaction (Model I) with one only allowing for main effects (Model II)--that is testing AB eliminating A and B. In a two factor complete factorial experiment this is the usual test of the two-way interaction which is routinely employed. If we are able to reject the hypothesis of null interaction effects our usual procedure would be to stop at this point with an interaction model. If, however, we are unable to rej[illegible] this hypothesis (i.e., conclude that interaction effects are non-significant) we would wish to proceed with tests of main effects.

When we allow the possibility of both an A and B effect in the population we are specifying a series of tests involving model II. Thus, to test either effect we must test it in that model, implying an alternative model in which it is absent. To test for an A effect we compare model II to model III, while to test for a B effect we compare model II to model IV. In each of these tests we are allowing for the possible existence of the effect not being tested. In testing A we are asking the question "given the possible existence of B in our model, do we need A?" This is the meaning of the term "A eliminating B".

Our judgment as to which model to accept is based upon the relative magnitudes of the sum of squared errors produced by the competing models and the F test gives a method for testing whether the models differ in this respect. This procedure is always correct, in either the orthogonal or non-orthogonal case. In the orthogonal case it will produce results identical to those produced by the ordinary computational methods.

Different tests of A and B effects may be appropriately obtained by beginning with different model assumptions. If we assume that there is no B effect, model IV is an appropriate model and we would compare it to model V in order to test the existence of an A effect in model IV (i.e., without regard to the existence of a B effect). This test of "A ignoring B" is not a proper test unless model IV is the correct model, i.e., unless there is no B effect. Similarly we may test B ignoring A by comparing model III against model V, but here the test is proper only if model III is appropriate, i.e., there is no A effect. In the case of an orthogonal design these tests will give us the same results as those tests involving model II, but while the results are computationally the same (due to independence of the estimates of the parameters involved) they are not logically the same in terms of comparing the same models.

An Example

Let us now apply this method to the data of Table 1 using the MANOVA computer program (Cramer, 1967). We may summarize all the relevant statistical tests in the following ANOVA Table (Table 5):

Table 5

ANOVA Tables for Complete Analysis of Data in Table 1

Source	df	SS	MS	F	p
AB	1	0.00	0.00	0.00	1.000
A eliminating B	1	370.37	370.37	1.646	.205
B eliminating A	1	0.00	0.00	0.00	1.000
A ignoring B	1	1349.99	1349.99	5.999	.017
B ignoring A	1	979.63	979.63	4.353	.042
Within Cells	50	11250.00	225,00		

It may be clearly seen that there is no evidence for an interaction; however, the small numbers of observations in two of the cells makes the power of this test rather low. Tests of A eliminating B and B eliminating A are clearly non-significant, while the tests of A ignoring B and B ignoring A, given previously, are both significant. All five of these statistical tests must be considered in order to draw proper conclusions about the population means. The tests of A eliminating B and B eliminating A do not provide us with any evidence regarding the existence of either A or B effects (although they clearly imply that both effects are not necessary jointly), while the tests of A ignoring B and B ignoring A separately provide us with evidence for either effect, depending upon which test we consider. Keeping in mind the models which are compared, the "eliminating" tests tell us that we have no evidence for one effect in addition to the other. We must conclude then from this statistical analysis that there must be some effect, either an A or B effect, but we cannot tell which, and there is, clearly, no evidence to suppose that both exist. This is in line with the previous conclusion obtained by informal arguments earlier. It should be noted that because of the substantially disproportionate numbers of observations in cells, the power of the eliminating tests is rather low and the effects are highly confounded. Indeed, this example approaches closely the completely confounded case in which all observations would be in the A_1B_1 and A_2B_2 cells. In the completely confounded case, the one degree of freedom between cells could be attributed to either an A effect or a B effect with no possibility of deciding between them.

Interpretation of Results

The patterns of possible results from the analysis of a two factor design with no interaction are given in Table 6. Pattern 1 indicates that

Table 6

Pattern of Results--Two-way Factorial without Interaction

	Pattern						
Test	1	2	3	4	5	6	7
A eliminating B	s	s	ns	ns	ns	ns	ns
B eliminating A	s	ns	s	ns	ns	ns	ns
A ignoring B	x	x	x	ns	s	s	ns
B ignoring A	x	x	x	ns	s	ns	s

s=significant ns=nonsignificant x=irrelevant

A and B are both needed in the model since, given the presence of one, the other is still significant. Patterns 2 and 3 both illustrate cases for which a second main effect is not needed given the inclusion of the other, but the significant effect must be included (i.e., From Pattern 2 we would retain the A effect, from Pattern 3, the B effect). Pattern 4 is the case for which no main effects are included in the final model. These constitute the standard, easy to interpret cases and are the only cases which may arise from an orthogonal design. The remaining patterns are unique to the non-orthogonal case. Pattern 5 is the seriously confounded situation presented earlier in which only one effect need be included in the final model, but due to confounding the choice of which effect to retain is indeterminant. Patterns 6 and 7 occur only in situations in which there is very serious confounding in the design.

The significant main effect should be included in the final model. In these circumstances it is particularly important to ask why such a seriously confounded design was produced and to carefully attend to the implication this has to the phenomena being investigated.

Recommended procedure for a two-way non-orthogonal design

On the basis of the material developed to this point we suggest the following procedure be employed in the analysis of a non-orthogonal two factor design. It should be emphasized that this procedure is for the logical flow of decisions and conclusions which are made in such an analysis, but does not dictate the actual order in which the computations need be performed. Indeed, in most of the standard computer programs available for such an analysis (e.g., MANOVA; Cramer, 1967) the required tests would be produced in a rather different order.

However, once the results of all required tests are available, we would suggest proceeding as follows:

A. Begin with the full model including main and interaction effects.

B. Test for a significant interaction (AB eliminating both A and B), if this test is significant no tests of main effects are appropriate; however, one might wish to test certain contrasts in the cell means to aid in interpretation of the results. If the test is non-significant eliminate the γ_{ij} terms from the model and proceed to step C for tests of main effect.

C. Test A eliminating B and B eliminating A

1. if both tests are significant adopt the model $Y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon$
2. if only one of the two tests is significant adopt the model $Y_{ij}=\mu+\alpha_i+\varepsilon$ (if A eliminating B is the significant one) or $Y_{ij}=\mu+\beta_j+\varepsilon$ (if B eliminating A is the significant one).
3. if neither is significant proceed to D.

D. Test A ignoring B and B ignoring A

1. if both are significant retain either α_i or β_j, but not both in the final model--the choice is indeterminant. In this case additional experimental evidence will usually have to be obtained before much could be said about the meaning of the experiment.

2. if only one of the two tests is significant, the significant effect should be retained, but the cautions referred to in the discussions of patterns 6 and 7 should be diligently adhered to.
3. if neither test is significant no main effects should be included in the first model, i.e., adopt the model $Y_{ij}=\mu+\epsilon$.

Extension to Higher Order Designs

As a non-orthogonal design becomes more complex through the inclusion of additional factors the proper analysis becomes far more tedious although the basic logical structure remains the same. In all cases we are attempting to find the simplest model which adequately fits the data by comparing competing models. As the number of factors increases the total number of potential tests (model comparisons) increases very rapidly. For a q-factor design the total number of potential tests is given by

$$\sum_{i=1}^{q} {}_{q}C_{i}2^{[({}_{q}C_{i}-1)]}$$

where ${}_{q}C_{i}$ is the number of combinations of q things taken i at a time. In most cases, however, not all tests will be performed.

Because of certain symmetries which exist in the three factor case, the extension of the two factor procedure to higher order designs is most easily seen through the analysis of the three factor design. In general the process begins by determining if the triple order interaction is necessary. If it is not, one proceeds to determine how many and which second order interaction, if any, are necessary and finally, in the absence of second order interaction, how many and which main effects are necessary in the model.

As a general point it should be noted that when a second order interaction is included in a model (say the $\beta\gamma$ term), the main effects implied by that term (in this case β and γ) will be also included; the other main effect terms (in this case α) may or may not be needed in the model. To determine if other main effects should be included requires a separate set of tests.

Procedure for a three factor non-orthogonal design.

The process begins by tentatively adopting the full model, $Y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$ and then eliminating unnecessary terms. First one would test the triple order interaction, ABC, eliminating all second order interactions and main effects, i.e., asking the question--given the lower order effects do we need the triple order interaction? If the test of the ABC interaction is significant, one would accept the full model and proceed, if desired, to test specific contrasts in cell means to aid with interpretation. If on the other hand the triple order interaction is non-significant indicating that the effect is not required in the model, given the possibility of lower order effects, one would drop the $\alpha\beta\gamma_{ijk}$ term from the model and would proceed to investigate the second order interaction terms in order to determine how many and which terms to include in the model.

At this point in our discussion, however, we shall consider the procedure for main effects rather than second order interactions. We do this because some of the concepts carry over directly from the two factor design and, given certain symmetries in the three factor design, it is possible to then directly apply these concepts to tests of interaction terms. We must emphasize that in the actual use of the process, tests of second order interaction would *always* preceed tests of main effects.

On notation

In order to simplify the naming of various tests (model comparisons) in the discussion to follow the following notational scheme will be used

(1) the symbol | will be used to indicate eliminating

(2) the absence of a term to the right of the | symbol of the same order as the term on the left of the | implies that term is ignored

(3) it is assumed that *all* lower order terms are eliminated from higher order terms.

Thus, for a three factor design with factor A, B, C

A|B,C implies the test of A eliminating B and C

A|B implies the test of A eliminating B and ignoring C

A implies the test of A ignoring B and C.

AB|AC,BC implies the test of AB eliminating AC, BC, A, B, and C

while AB implies the test of AB ignoring AC and BC but eliminating A, B, and C.

Tests of Main Effects

In testing for main effects we are trying to determine how many effects must be included in the model and which ones they are. The only circumstance under which it would be necessary to include all of the main effects is when each main effect is significant eliminating the other two, i.e., when the tests A|B,C; B|A,C; and C|A,B are all significant. If only two of the three tests of main effects eliminating both of the others are significant, the two significant effects would be retained while the third would be deleted from the model. Thus if all three of the tests or if two of the three tests are significant our conclusions are quite direct--retain the significant effects.

When, however, only one or none of the three tests is significant the situation is somewhat more complex. If only one of the main effect terms eliminating the other two is significant, say A|B,C, the significant term should clearly be retained; however, it may be desirable to retain one of the other two effects. Since we have already decided to keep the A effect we need ask do we need <u>either</u> the B or the C effect given the A effect, i.e., to test B|A and C|A. If neither of these tests is significant then clearly neither effect needs be present given the A effect in the model. If one of the two is significant, say C|A, that term, C, should be included in the final model along with the A term. Should, however, both be significant, we are in an ambiguous situation. Previous tests have indicated that all three effects are not needed in the model and that the A effect must be in the model, therefore our choice between B and C is completely indeterminant.

The potentially most complicated situation obtains when none of the three "doubly eliminating" tests are significant. It is still possible that one or two effects should be included in the model. In the two factor design, we reasoned that the significance of both A|B and B|A indicated that both A and B should be included. In the three factor design there are three such pairs of tests involving A and B, A and C, and B and C (i.e., A|B and B|A; A|C and C|A; and B|C and C|B). The joint significance of any one of these pairs of tests indicate the need to include the relevant pairs of effects, but only two such effects may be included, our previous tests having excluded the possibility of all three effects being included in the model. If more than one pair of these tests shows significance we are uncertain as to which pair of effects to include. This is analogous to the two factor case where we were uncertain as to which of the two main effects to include. Should no pair of effects be significant we are then left with the possibility of including only a single effect in the model. Thus if any one effect were to appear significant (e.g., if the tests of A|B or A|C were significant) we would include it in the model. Should none of the "single eliminating" tests be significant we would then examine the "doubly ignoring" tests, A, B, and C as these may still indicate the necessity of including a single main effect. If none of these tests are significant we would conclude that no main effects were necessary and would be left with the model $Y_{ijk}=\mu+\epsilon_{ijk}$. If but one of these tests is significant, that effect would be included in the final model. If two or more of the "doubly ignoring" tests are significant we are again in an indeterminant situation and may arbitrarily choose one of the significant effects for the final model, but the choice is completely arbitrary.

Application to two-way interactions

The application of the "main effect procedure" to two-way interaction is straight-forward if we but note the following symmetry which exists in the three factor case. Since there are three two-way interactions and three main effects in a three factor model, the pattern of tests for main effects and for two-way interactions are exactly the same. Corresponding to tests of main effects A, B, and C there are tests of interactions AB, AC, and BC. For every main effect test, say A|B,C, there is a corresponding test AB|AC,BC. Hence, the the above procedure is first applied to the three two-way interactions eliminating all main effects and other two-way interactions, i.e., AB|AC,BC, AC|AB, BC, and BC|AB,AC, and would then be followed with parallel tests as needed. Should the conclusion be that there are no interactions, the procedure would then be applied to the main effects. If there are significant interactions, the factors involved should be also included as main effects, as noted earlier. Should only one two-way interaction be included, the question of retaining the uninvolved main effect should be considered. To do this the test of that effect eliminating the other two main effects and the significant interaction should be performed, e.g., if it were the BC interaction that were significant one should perform the test A|B,C;BC in order to determine if the A effect should be included in addition to the B, C, and BC effects.

Some additional comments

The methods discussed for both the two and three factor cases have proceded on the assumption that there is no a priori preference for explaining the data in terms of one factor above any others. Such a preference may exist in designs such as randomized blocks where we would customarily not even consider the test of treatments ignoring blocks; we assume that there are block effects and are willing to consider the presence of treatment effects only if the test of treatment eliminating blocks is significant. Similar considerations may apply in a wide variety of cases and may simplify the process discussed here.

Another consideration is the number of tests involved in the complete procedure. Some of these tests will be highly correlated and some will be independent depending upon the degree and pattern of non-orthogonality. The extreme is illustrated by the two factor orthogonal case in which the tests A and B are independent while A|B and A are identical. In the case of lack of knowledge of likely effects one may perform preliminary combined tests such as a test of pooled interaction prior to doing individual tests. This would have to be moderated, however, by any knowledge which would, a priori, suggest the existence of specific effects.

Overall, Spiegel, and Cohen (1975) have considered some of the problems discussed above and have arrived at very different and demonstrably much more limited conclusions. Since this is so relevant to balancing, we will indicate the serious flaws in their "proper" generalization of orthogonal ANOVA

An Analysis of the Recommendations Presented by Overall, Spiegel, and Cohen

Overall and Spiegel (1969) considered three methods of analysis in nonorthogonal ANOVA without favoring any one as being the appropriate one. Overall, Spiegel, and Cohen (1975) then argued that one of the three methods is indeed the only proper one to use. In describing how they arrived at this conclusion, they note that the strategy that "appeared most often to be recommended in applied statistics texts involves basically a 'main effects' model with tests for interaction effects included as a safeguard against departures from additivity (Rao, 1965; Snedecor & Cochran, 1967; Winer, 1971). The analysis proposed by Appelbaum and Cramer (1974) follows this logic" (p. 184). The argument against this approach, as developed by Overall et al., rests upon a single principle which we believe to be correct and proper, and a single procedure which is easily demonstated to be erroneous.

The principle is "that the method for the analysis of variance of data from nonorthogonal designs should estimate the same parameters and test the same hypotheses as can otherwise be estimated and tested in a balanced analysis of variance and experimental design involving the same factors" (p. 184). This is consistent with our views since in our 1974 paper we said, "Having decided to employ the method of least squares...one is left only with the selection of possible models and model comparisons. The models selected are logically independent of the observed numbers of observations per cell" (p. 336). The key point which Overall et al. ignore is the choice of model. Given a model, we would argue that our methods test the same hypotheses and estimate the same parameters whether there are equal numbers of observations or not. In the absence of a model we believe it to be meaningless to talk of estimating parameters, much less testing them.

The procedure Overall et al. propose for verifying that a particular method satisfies the above criterion is "to generate data for orthogonal and nonorthogonal designs involving exactly the same α_i, β_j, and $(\alpha\beta)_{ij}$ and then to determine which method of analysis yields the same parameter estimates in the orthogonal and nonorthogonal cases." This procedure is ill-defined since it does not state how such data should be generated. If the example presented by Overall et al. is meant to make the procedure precise, it is clear that their procedure is patently inappropriate. Overall et al. present data arranged in a 3x3 ANOVA with thre observations per cell and then duplicate the observations in certain cells to make the design nonorthogonal. They state that "the reader will appreciate that duplication of certain scores does not invalidate the analysis of variance" (p. 184). Quite the contrary, it does invalidate the analysis of variance since the observations are clearly not independent in the various cells. Furthermore if they claim (as they appear to) that the addition of observations should not change the estimates of the parameters, it must be the case that the method ignores, in generating estimates, any information in the additional observations. How can a method that ignores such information be a good method?

We have analyzed the data given by Overall et al. and we suggest that even if one ignores the question of independence and follows the procedures we have advocated he will not perform any tests of "main effects" for the simple reason that there is a significant interaction in the data which they present. (A detailed discussion of the problems involved in testing and estimating "main effects" in the presence of an interaction follows.) Analyzing their data with the additional observations, we obtain an F value of 6.4 for the interaction which is significant beyond the .001 level. Given this result we would probably wish to look at A effects for given levels of B, or B effects for given levels of A, or possibly individual interaction contrasts. We doubt that we would have any interest whatsoever in any of the parameters that Overall et al. obtain or in any of the main effect tests they perform. Indeed, we have made in our earlier specific recommendations as follows:

1. Begin with the full model including main effects and interactions effects.

2. Test for a significant interaction; if this test is significant no tests of main effects are appropriate; however, one may wish to test certain contrasts in the cell means to aid in interpretation of the results.

Procedures for the Case of Significant Interaction

Since, in our previous work, we were not specific about what we would do in the case of a significant interaction, it may be useful to consider our interpretation for this example. The cell means and numbers of observations are as shown in Table 7. Our standard ANOVA for an interactive model gives us an estimated standard deviation of 1.93 based on the within cells sum of squares. The marginal means shown are the unweighted means of the cell means for rows and columns. The significant interaction ($F = 6.4$, $p<.001$) tells us that there are effects of both A and B, but that the A effects are different for different levels of B just as the B effects are different for different levels of A. It seems clear from examination that the interaction is due primarily to the value 11.7 in cell 13. If we delete that cell we can obtain the test of that portion of interaction remaining with three degrees of freedom rather than four.[2] On reanalysis, with the 13 cell deleted, we find that the interaction is no longer significant ($p=.27$), strengthening the belief that this one cell is responsible for the significant interaction. The test of A eliminating B is highly significant ($p<.001$) while the test of B eliminating A is marginal ($p=.10$). It appears then that if cell 13 is dropped there is definitive evidence only for an A effect.

[2] Analysis of variance programs such as MANOVA (Cramer, 1967) allow for the complete deletion of specified cells making such an analysis a simple matter.

As an alternative or supplementary analysis, we have analyzed the simple effects of A for each level of B and the simple effects of B for each level of A. These analyses also confirm what inspection of Table 7 suggests; the simple A effects for levels one, two, and three of B are highly significant (p=.009, .001, .001). The simple B effects for levels one and three of A are significant (p=.001, .03); the simple effect of B for level two of A is not significant (p=.81).

On Main Effects, Marginal Effects, and Interaction

Additional insight into the nature of this problem can be gained through a more careful consideration of the problems of testing and estimating "main effects" in the presence of interaction. At this point it is necessary to introduce a basic logical distinction between two concepts which have, unfortunately, come to be held as virtually synonymous--a main effect and a marginal effect. By a main effect we mean the effect of a particular experimental treatment or state of nature which is the common and consistent effect of that treatment or state of nature irrespective of what other treatments or states of nature it is combined with. By a marginal effect we mean simply the average effect of the experimental treatment (state of nature) averaged, in some sense, over all occurrences of that treatment. These two concepts are equivalent only in the noninteractive model. In the case of a model in which there is an interaction, the two concepts are quite distinct; in fact, under the interactive model, the concept of a main effect does not apply, for an interaction implies that there is no consistent effect of the treatment,

Table 7

Means and Numbers of Observations for Data from Overall, Spiegel, and Cohen

	B_1	B_2	B_3	
A_1	6.0 6	5.7 3	11.7 3	7.8
A_2	6.3 3	6.0 6	5.3 3	5.9
A_3	10.3 3	13.0 6	10.0 6	11.1

Table 8

Illustration of Marginal Means for Interactive Model

	B_1	B_2	w=.5	w=1	w=0
A_1	10	20	15	10	20
A_2	20	10	15	20	10

Table 9

Illustration of Marginal Means for Non-interactive Model

	B_1	B_2	w=.5	w=1	w=0
A_1	10	20	15	10	20
A_2	30	40	35	30	40

but rather that one must consider a treatment in combination with some other treatment(s) in order to assess its effect. This distinction can also be seen through the concept of a simple row (or column) effect which is commonly defined as the difference between a cell mean and its corresponding row (or column) mean. If, ıor a given factor, the simple effect of the several treatments sнould be identical for all levels of other factors, this constant simple effect is the main effect.

Given then that one is operating with a model which contains an interaction term it is, at best, misleading to speak of main effects, for one is considering marginal effects. These marginal effects will be averages of cell means across rows or columns of the design. There is no particular reason for using a simple average rather than a weighted average. If the model is truly interactive the weights used will have a substantial effect on the marginal effects. Suppose, for example, the cell means are as shown in Table 8 for a two by two ANOVA. If we define a marginal A mean as

$$\hat{Y}_{i.} = w\,\hat{Y}_{i1} + (1-w)\,\hat{Y}_{i2}$$

we find that the difference in marginal means for A will be 0, -10, or 10, depending on whether w is .5, 1, or 0. For the data from a noninteractive model shown in Table 9, we find that the difference in marginal means is 20 regardless of what the weights are.

The tests of main effects proposed by Overall et al. in Method I are in fact tests of equally weighted marginal means for an interactive model. It can also be shown that these tests are equivalent to the method of unweighted squares of means proposed by Yates (1934) and discussed by Bancroft (1968). These are tests of the equivalence of row or column marginal means

$$\mu_{i.} = \sum_j \mu_{ij}/b$$

and

$$\mu_{.j} = \sum_i \mu_{ij}/a$$

These particular marginal means are but one of many possible sets of marginal means which can be constructed and it is by no means clear that this is the most desirable set to test in any particular situation (see Appelbaum & Cramer, 1975).

We believe then that the above analyses reveal essentially everything there is in the data. As we have indicated, the tests of main effects recommended by Overall et al. are equivalent to the tests of equality of marginal means as we have defined them. We do not find these tests very interesting since the marginal means represent only average effects for rows and columns, while the significant interaction tells us that these average effects are different from the actual effects for each row and column. The marginal A effect is significant (p=.001); the marginal B effect is not (p=.23).

We would argue then that the example presented by Overall et al. does not bear on the validity of the methods we have advocated, for the simple reason that there is an interaction present. Furthermore it seems to us that their analysis of "main effects" is not directed to the questions that psychologists will typically wish to address. We could of course modify their example so that the interaction is nonsignificant. Then, as we have noted, it would violate the assumptions of independence. Their procedure is simply not valid in principle.

Estimation and the Overall et al. Criterion

Overall et al. have erred in assuming that if two methods estimate the same parameters, they must yield the same estimates. This is obviously false. To estimate a population mean we could use a sample mean or simply use the first observation, discarding the others. Both the sample mean and the first observation are unbiased estimators of the population mean, but they will, in general, yield different estimates. The sample mean is better since it will be closer to the population mean on the average. This precision of estimates is the crucial distinction between the methods Overall et al. advocate and the methods we advocate.

In our 1974 paper the topic of estimation in the nonorthogonal ANOVA did not appear since we did not believe that there was any disagreement as to what was appropriate. We now feel that this topic does require some attention.

The estimation problem is easily and completely solved once one decides upon the model which one believes applies to the real world. The usual role of significance testing is to determine, based upon the data of the real world, which model is the most reasonable one from among a set of competing models. Having made a decision as to which model obtains, one may then proceed to estimate the parameters of the model--but estimation may occur only in the context of a particular model.

Let us now consider one possible model--the two factor interactive model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon \qquad (1)$$

It is a trivial matter to obtain a set of least squares estimates of the parameters of this model. We say "a set" because there are infinitely many sets which are equivalent in the sense that they will yield identical predicted values $\hat{Y}_{ij}$. It is, however, a standard practice to impose additional constraints upon the model in order to obtain a unique set of estimates. The purpose of the constraining system, however, is solely computational convenience. It is obvious that the very best we can do in this model is to predict the cell means exactly, since there are no parameters which are unique to any single observation. Any two models which predict the cells means exactly must be equivalent. It is also a consequence of the mathematics of the system that any model which has as many independent parameters as cells must predict the cell mean exactly.

There exist infinitely many constraining systems which may be applied to the full interactive model in order to produce the computational determinacy desired. The simplest of these is

$$\mu = \alpha_i = \beta_j = 0 \qquad \text{(for all i and j)}$$

leaving the model

$$Y_{ij} = \gamma_{ij} + \varepsilon \qquad (2)$$

In this case the Least Squares estimates of the γ_{ij} are simply the observed cell means, $\bar{Y}_{ij}$.

The more usual (conventional) constraining system, however, is

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \qquad (3)$$

If the design has a levels of factor A and b levels of factor B, there are then (after the imposition of the constraints) $1 + (\underline{a}-1) + (\underline{b}-1) + (\underline{a}-1)(\underline{b}-1) = \underline{ab}$ independent parameters which is also the number of independent parameters in (2). This equals the number of cells in the design, and it then follows that the model constrained in this way must be equivalent to that in (2). For those familiar with the matrix approach to the analysis of variance this result is easily seen from the fact that the model matrix for this constrained design must have ab columns.

It is also a rather trivial matter to directly write the least squares estimates of the parameters of the interactive model constrained by (3). They are

$$\begin{aligned}\hat{\mu} &= \bar{Y}_{..}\\ \hat{\alpha}_i &= \bar{Y}_{i.} - \bar{Y}_{..}\\ \hat{\beta}_j &= \bar{Y}_{.j} - \bar{Y}_{..}\\ \hat{\gamma}_{ij} &= \bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\end{aligned} \qquad (4)$$

where $\bar{Y}_{..}$ is the unweighted average of the cell means while $\bar{Y}_{i.}$ and $\bar{Y}_{.j}$ are the unweighted averages of the cell means for row i and column j, respectively. Substituting these estimates into (1) gives

$$\begin{aligned}\hat{Y}_{ij} &= \mu + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}\\ &= \bar{Y}_{ij}\end{aligned}$$

again showing the equivalence of (1) and (2).

We thus see that estimation in the interactive model is rather trivial, with the estimates of the parameters being simple linear functions of the observed cell means and free of the n_{ij}.

In point of fact there is really no gain in talking of estimating parameters in (1) since it is equivalent to (2) which is a cell means model requiring no constraints. We have a x b populations (one per cell) and the only parameters of interest are their means and their common variance.

The situation, in general, is not nearly so simple when there is no interaction, that is, when estimation proceeds within the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon \qquad (5)$$

In general we would have to solve a set of simultaneous least squares equations in order to obtain estimates of parameters in (5). An exception occurs, however, in the orthogonal case in which the estimates of μ, α_i, and β_j have the form as in (4). In the more general nonorthogonal case, there will again be infinitely many solutions to the unconstrained least squares equations although estimates of $\mu + \alpha_i$ and $\mu + \beta_j$ will be unique.

An interesting result of least square estimation in (5) is that the estimates obtained for μ, α, and β from model (1) yield unbiased estimates of Y_{ij} in (5), but the estimates are less precise, that is, they have larger variances than the estimates obtained from (5). For this reason it will be desirable to estimate the parameters of (5) when we have accepted (5) as the true model rather than use the estimates from model (1).

We would agree that the procedures advocated by Overall et al. test the same hypotheses in both the orthogonal and nonorthogonal case; further, we agree that they are valid tests of certain hypotheses, but we doubt that they are hypotheses of particular interest in either the orthogonal or nonorthogonal case. We believe that an informed statistical analyst would not perform a test of main effects in the presence of a significant interaction in the orthogonal case; why then in the nonorthogonal case?

The Case of Nonsignificant Interaction

Of course, we can only know in a probabilistic sense if there is truly an interaction present in nature. We must in the final analysis rely on the results of statistical tests to direct us to reasonable models upon which to base our estimation procedure. This then leads us to ask what behavior is appropriate when the data dictate an interaction free model and to consider the consequences of such behavior. There are, in this respect, only three cases which need concern us; in all three we will assume that the statistical test of interaction is nonsignificant.

Case 1: No interaction in the population

The first case we shall consider is the case when indeed there is, in nature, no interaction present. No empirical demonstration is needed to verify that if one has the form of the true linear model, the least squares estimates of the parameters in that model will be the best unbiased linear estimates. Furthermore, it is completely obvious that if one fits an interactive model when there is in fact no interaction, one will obtain unbiased estimates which will not be minimum variance. For this reason it is a mistake to include worthless effects in an ANOVA model, just as it would be to include worthless variables in a regression problem. The additional sampling error causes the main effect parameters and the estimated cell means to have larger standard errors than would the estimates from a main effects model. This point has been noted in a regression context by Walls and Weeks (1969) and is exactly what would occur if Overall and Spiegel's Method I were applied in this case. The increase in sampling error may be quite substantial and will result in less powerful tests of main effects.

Case 2: Small but nonsignificant interaction effects

The second case is the situation in which there is a true interaction but its magnitude is too small to be detected by a conventional test of interaction. We have previously argued that the main effect parameters are not meaningful for the interactive model, but that the predicted cell means are. The predicted cell means will have a smaller variance when estimated in the main effects model than when estimated in the interactive model since the variance depends only upon the design matrix (X in the usual matrix approach to ANOVA) and the variance of the dependent variable. The predicted cell means will, however, be biased in this case. Since we can no longer speak of minimum variance unbiased estimators, it then becomes the mean square error which is relevant for comparison. We must add the mean squared bias (which will be a function of the magnitude of the small but nonzero interaction terms) to the variance to obtain the mean square error. This term will be small if the interactive effects are small as would be the situation under Case 2. Operating under Case 2, we will still be estimating the same parameters and testing the same effects in both the orthogonal and nonorthogonal cases, but we will simply be estimating and testing with a small amount of bias. We will gain substantially in that the estimates will be more precise and the tests will be more powerful than if we followed Overall and Spiegel's Method I which is based upon the interactive model.

To see the difference, let us compare the variances of the estimated parameters and estimated cell means for the data used by Overall et al. In Table 10 we have computed the variances of the estimated main effect parameters and the predicted cell means for both the main effects model (our procedure) and the interactive model (Overall and Spiegel's Method I). If X is the matrix of independent variables, the variance-covariance matrix of the estimated parameters is $(X'X)^{-1}\sigma^2$ while the variance-covariance matrix of the predicted cell means is $X\ (X'X)^{-1}X'\sigma^2$. The variances are on the diagonals of these matrices and do not depend upon which model is correct in nature. Since σ^2 serves only as a scale factor, we have assumed in Table 10 that it is equal to one. We see that the estimated parameters of the main effects model have slightly smaller variances than those of the interactive model, while the corresponding predicted cell means have substantially smaller variances when estimated from the main effects model. The effect on the predicted cell means is particularly marked for the cells with a small number of observations, since the variance of a sample mean (the predicted value for an interactive model) is simply σ^2/n.

<u>Case 3</u>: A large interaction which is not detected

The third case covers the situation where a large interaction

Table 10

Variances of Parameter Estimates and Predicted Cell Means for 3x5 Factorial Design with Unequal Numbers of Observations Assuming $\sigma^2 = 1$.

	Main Effects Model	Interactive Model
Parameters Estimated		
μ	.161	.169
α_1	.241	.244
α_2	.235	.244
β_1	.241	.244
β_2	.224	.231
Estimated Cell Means		
$\hat{Y}_{11}$	.116	.167
$\hat{Y}_{12}$	.151	.333
$\hat{Y}_{13}$	.158	.333
$\hat{Y}_{21}$	.158	.333
$\hat{Y}_{22}$	.112	.167
$\hat{Y}_{23}$	.154	.333
$\hat{Y}_{31}$	.151	.333
$\hat{Y}_{32}$	.109	.167
$\hat{Y}_{33}$	.112	.167

is somehow not detected by the interaction test. In this situation the reverse of Case II will occur and the mean square errors of the estimated parameters and cell means will be small for the interactive model. The probability of this third case occurring is, however, remote, for if the magnitude of the interaction effects is large and if the sample size is reasonable the power of the interaction test is quite large.

We have shown above that the significance testing procedures which we have previously recommended for the nonorthogonal ANOVA are consistent with the basic principle advocated by Overall et al.; namely, that in the nonorthogonal case one should estimate the same parameters and test the same hypotheses that one would estimate and test if there were equal numbers of observations in the cells. Indeed, that principle is implicit in our original paper. We have pointed out that our method of fitting a series of main effect models (in the absence of interaction) is <u>not</u> the same as their Method II. We have further shown that their method for achieving the stated goal is incorrect and, if routinely applied, will not lead to optimal tests or estimates. In discussing the relationship between estimates and hypothesis testing we believe that we have made clear the reasons for preferring our procedure since it leads to more powerful tests and more precise estimates.

It must be recalled that the issue of how to estimate effects and how to test hypotheses are rather distinct. The methods discussed by Overall et al. and by us are methods for testing hypotheses and not for the estimation of effects. This distinction is not one which we uniquely make. Bock (1975), for instance, regards these as distinct processes. He begins with some initial model, performs tests of significance to determine if a simpler model is appropriate, and then estimates the parameters in the simplest reasonable model. We have seen no evidence which suggests that the methods advocated by Overall et al. are preferable. We continue to maintain, along with Rao and others, that one should test main effects, assuming no interaction to be present, when this is what is suggested by the data at hand.

Chapter IV: A Comparison of Balancing and Other Methods of Adjustment

Several alternative methods are available for adjusting for group differences in a dependent variable when the groups are not randomly constituted and thus may exhibit systematic differences on interfering variables that are related to the dependent variable. The best known of these methods is analysis of covariance. Other methods, based upon somewhat different assumptions, include direct and indirect standardization and balancing.

Analysis of Covariance

Analysis of covariance (see, e.g., Elashoff, 1969; Tatsuoka, 1971) assumes that, in the population of interest, the i'th person in the j'th group has a score Y_{ij} on the dependent variable that can be expressed as

$$Y_{ij} = \mu_j + \beta\ (X_{ij} - \bar{X}) + e_{ij}$$

(where $\mu_j = \mu + \alpha_j$). In this notation μ_j is the adjusted population mean on the dependent variable for the j'th group; β is the within-group regression coefficient; X_{ij} is the score on the interfering variable (the variable for which adjustment is made) for the i'th person in the j'th group; $\bar{X}$ is the mean score of the observations over all groups on the interfering variable; and, e_{ij} is an error term for the i'th person in the j'th group. The mean score on the dependent variable for the j'th group can be expressed as

$$\bar{Y}_j = \mu_j + \beta\ (\bar{X}_j - \bar{X}) + \bar{e}_j \quad ,$$

where $\bar{X}_j$ is the mean observed score on the interfering variable for group j and $\bar{e}_j$ is the mean error term for group j.

For more than one interfering variable, a model of the following form is used:

$$\bar{Y}_j = \mu_j + \sum_m \beta^{(m)}\ (\bar{X}_j^{(m)} - \bar{X}^{(m)}) + \bar{e}_j$$

where $\beta^{(m)}$ is the within-group multiple regression coefficient for the m'th interfering variable; $\bar{X}_j^{(m)}$ is the mean score for the j'th group on the m'th interfering variable; and, $\bar{X}^{(m)}$ is the mean score over all groups on the m'th interfering variable.

Least squares estimates for the parameters of the model are obtainable and parameter values can be replaced by these estimates to obtain an estimate of the adjusted mean score for the j'th group,

$$\hat{\mu}_j = \bar{Y}_j - \sum_m \hat{\beta}^{(m)} (\bar{X}_j^{(m)} - \bar{X}^{(m)}) \quad .$$

Balancing, Direct Standardization, and Indirect Standardization

With one interfering variable, the model used in balancing, direct standardization, and indirect standardization is the additive analysis of variance model. This model assumes that in the population the i'th person in the j'th group with a score at the k'th level of an interfering variable has a score Y_{ijk} that can be expressed as

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + e_{ijk} \quad ,$$

or

$$Y_{ijk} = \mu_j + \gamma_k + e_{ijk} \quad .$$

The mean score for persons in the j'th group and the k'th level of the interfering variable then can be expressed as

$$\bar{Y}_{jk} = \mu_j + \gamma_k + \bar{e}_{jk} \quad .$$

In this notation μ_j is the adjusted mean (in the population) on the dependent variable for the j'th group and γ_k is an effect associated with the k'th level of the interfering variable.

With more than one interfering variable, balancing still may be employed, based upon the additive analysis of variance model. Direct standardization and indirect standardization usually are defined only for one interfering variable. However, each can be generalized to accomodate more than one interfering variable. The generalized model for either direct or indirect standardization also allows for all possible interactions among the interfering variables. For example, for three interfering variables, balancing employs a model of the form

$$\bar{Y}_{jk\ell m} = \mu_j + \gamma_k + \delta_\ell + \tau_m + \bar{e}_{jk\ell m} \quad ;$$

generalized direct and indirect standardization are not restricted to an additive model and may use a model of the form

$$\bar{Y}_{jk\ell m} = \mu_j + \gamma_k + \delta_\ell + \tau_m + \gamma\delta_{k\ell} + \gamma\tau_{km} + \delta\tau_{\ell m} + \gamma\delta\tau_{k\ell m} + \bar{e}_{jk\ell m} \ .$$

More will be said about generalized direct and indirect standardization in a later section.

While the analysis of covariance model (of the previous section) can treat interfering variables either as continuous or discrete, these analysis of variance models must treat all interfering variables as discrete, since each is to be categorized by level. However, measurement on a continuous variable must always result in observed values on a discrete variable, since the measurement process yields a finite set of possible values while the number of possible values of a continuous variable is indefinitely large (see Jones, 1971). Thus, the operative distinction is that, with ANCOVA, an interfering variable may be measured with an indefinitely large number of score categories, while with balancing and standardization (direct and indirect), it is desirable that the number of score categories be limited. Results given by Cochran (1968a) suggest that only a slight loss of precision is associated with categorizing a continuous variable and then using standardization instead of using analysis of covariance with the original continuous variable as the covariate. Since we wish to compare ANCOVA, standardization (direct and indirect), and balancing, the remainder of this discussion assumes that variables are discrete, either because this was their original form or because they have been categorized.

Balancing -- The technique of balancing was developed for the National Assessment of Educational Progress in an attempt to present estimates of educational achievement that are relatively uncontaminated by interfering variables (see National Assessment of Educational Progress, 1973). Appelbaum and Cramer (1975) have shown that the estimates of parameters from balancing are the least squares estimates from an additive analysis of variance model, obtained by solving the normal equations. The primary estimates of interest are the estimates of the adjusted mean scores, the $\hat{\mu}_j$.

There is a systematic relationship between estimates obtained by ANCOVA and balancing when interfering variables are discrete. To understand this relationship, it must be realized that nonorthogonal ANOVA estimates can be obtained by using orthogonal polynomial contrasts for each interfering variable as covariates in the ANCOVA model, since both ANOVA and ANCOVA are part of the general linear model (e.g., Bock, 1975, chap. 5; Cohen, 1968). Analysis of covariance, however, uses only the linear trends of each factor as covariates, while balancing uses all trends of each factor as adjustment variates. Thus, estimates from the ANCOVA model are equivalent to those from a balancing model that assumes all trends other than the linear to be equal to zero.

The choice between using balancing or ANCOVA involves a trade between bias and variability of the estimates. Parameter estimates obtained from a model are unbiased only if that model is valid in the population (Draper and Smith, 1966). Analysis of covariance, unlike balancing, requires that all nonlinear trends be zero in the population if its estimates are to be unbiased. Thus, estimates from ANCOVA are at least as biased as estimates from balancing. On the other hand, since the parameters of the ANCOVA model are a subset of those of the balancing model, and since both models may be conceptualized as regression models, the variance of balanced estimates must be at least as large as the variance of estimates from ANCOVA (see Walls and Weeks, 1969, for the general regression case).

When it is certain that the relations between the interfering variables and the dependent variable are essentially linear, analysis of covariance is to be preferred to balancing, since estimates from both models are unbiased but those from the ANCOVA model are less variable. When relations are materially nonlinear, balancing is generally to be preferred to ANCOVA, but the magnitude of nonlinear trends and the sample size both should be considered. Estimates from balancing are less biased than those from analysis of covariance; the difference between the squared biases of the estimates from balancing and from ANCOVA depends upon the magnitude of nonlinear trends and is independent of sample size. However, the difference between the variances of the estimates from balancing and from ANCOVA is inversely proportional to sample size. With few observations, the difference between the variances is more likely to exceed the differences between the squared biases, in which case ANCOVA has the smaller mean-square error and on that basis ANCOVA is preferred to balancing. With a sufficient number of observations, however, the difference between the variances is unlikely to exceed the difference between the squared biases, in which case balancing, with a smaller mean-square error, is to be preferred. (An alternative under these conditions, not considered here, is a generalized ANCOVA, which adjusts for some subset of the nonlinear trends.)

Direct standardization -- Direct standardization has been used extensively by demographers, biostatisticians, and health researchers desiring to adjust for interfering variables in the comparison of group effects. Basic references for direct standardization are Fleiss (1973) and Kalton (1968). An example of the use of direct standardization is presented by Moses (1969) in connection with the National Halothane Study, where the desire was to assess the effects of halothane and other anesthetics on death rates, taking account of the differential patient characteristics associated with the use of various anesthetics.

With one interfering variable, direct standardization is based on the same model as balancing, but involves a different procedure for estimating parameters. The first step in this procedure is to estimate parameter differences of the form $\alpha_j - \alpha_{j'}$, where j and j' represent distinct groups. Direct standardization estimates this difference to be

$$\widehat{\alpha_j - \alpha_{j'}} = \frac{\sum_k w_k(\bar{Y}_{jk} - \bar{Y}_{j'k})}{\sum_k w_k}$$

where the w_k represent weights chosen by the experimenter.

Kalton (1968) has shown that, when adjusting for one interfering variable and comparing the means of two groups, a minimum variance estimator of this difference (assumed to be constant for all k) is obtained by choosing weights such that

$$w_k = \frac{n_{1k}n_{2k}}{n_{1k} + n_{2k}} \quad .$$

This derivation assumes equal variance within each cell of the design. While the assumption is unlikely to be valid for proportions, Kalton (1968, pp. 127-128) shows that for proportions this choice of weights usually is adequate and sometimes is preferable to the use of weights obtained assuming unequal cell variances. Thus, the difference between the means of the two groups at a particular level of the interfering variable is weighted inversely proportional to the variance of the difference between the means. Intuitively, when both groups are vell represented at a particular level of the interfering variable, the variance of the difference between the group means will be relatively small, and that level of the interfering variable will receive a relatively large weight in the estimate of the adjusted difference between the group means. Kalton states, "If the Y_{ijk} are normally distributed within subgroups, this

model is the usual fixed effects analysis of variance model, with

$$\widehat{\alpha_1 - \alpha_2} = \frac{\sum_k w_k(\bar{Y}_{1k} - \bar{Y}_{2k})}{\sum_k w_k}$$

estimating a main effect" (Kalton, 1968, p. 123; the notation is changed to correspond with that used here). Direct standardization with these minimum-variance-producing weights will yield the same estimates as balancing, i.e., the estimates from an additive analysis of variance model, with or without a normal error distribution.

Snedecor and Cochran (1967) show that, in an additive two-factor analysis of variance model where one factor has two levels, the estimated differential effect for that factor is given by

$$\widehat{\alpha_1 - \alpha_2} = \frac{\sum_k \frac{n_{1k}n_{2k}}{n_{1k} + n_{2k}} (\bar{Y}_{1k} - \bar{Y}_{2k})}{\sum_k \frac{n_{1k}n_{2k}}{n_{1k} + n_{2k}}} ,$$

confirming that direct standardization with this choice of weights does produce the same estimates as the least-squares estimation procedure of an additive analysis of variance model. When both factors have more than two levels, however, this choice of weights does not in general yield the same estimates as the least-squares estimation procedure used with the additive analysis of variance model. Thus, when more than two groups are to be compared, the weights presented by Kalton (1968) will not produce the same estimates as an additive analysis of variance procedure. Intuitively, this can be understood by noting that a comparison of two of the groups by direct standardization completely ignores all other groups when adjusting for the effect of the interfering variable. In contrast, the standard estimation method under the analysis of variance model uses all groups to estimate the effect of the interfering variable.

The remainder of the discussion of this weighting procedure assumes that only two groups are to be compared.

While direct standardization usually has been used only when one interfering variable is to be adjusted for, the procedure can be generalized to more than one interfering variable. For example, consider three interfering variables with K, L, and M levels. This design can be re-expressed as a design with one interfering variable with KxLxM levels. Direct standardization with minimum-variance-producing weights for this design will then yield the same estimates for adjusted group means as would the standard estimation methods for an analysis of variance model of the form

$$\bar{Y}_{jk\ell m} = \mu_j + \gamma_k + \delta_\ell + \tau_m + \gamma\delta_{k\ell} + \gamma\tau_{km} + \delta\tau_{\ell m} + \gamma\delta\tau_{k\ell m} + \bar{e}_{jk\ell m} .$$

With ·e interfering variable (and only two groups) balancing and direct standardization give the same estimates. With more than one interfering variable, the estimates from balancing and direct standardization generally will differ. Direct standardization requires that there be at least one observation for every combination of the interfering variables, so that

$$n_{1k\ell m} + n_{2k\ell m} \neq 0$$

for all k, ℓ, and m; otherwise, estimates cannot be obtained in this model, because division by zero would be required. In this study, involving five interfering variables, direct standardization is not employed because this requirement fails to be satisfied.

Another weighting procedure for direct standardization seems to be more widely used than that described by Kalton (1968). This procedure, described by Fleiss (1973), Cochran (1968a), and Moses (1969), uses weights of the form $w_k = \sum_j n_{jk} = n_k$, so that

$$\widehat{\alpha_j - \alpha_{j'}} = \frac{\sum_k n_k(\bar{Y}_{jk} - \bar{Y}_{j'k})}{\sum_k n_k} .$$

Both this weighting procedure and Kalton's (1968) yield unbiased estimates of the group effect in the analysis of variance model, but in general the variance of the estimate based upon this procedure is at least as large as the variance based upon Kalton's procedure. Intuitively, this procedure takes into account only the total number of observations for a level of the interfering variable; by neglecting how this total is distributed for the different groups, an unstable estimate of the mean for a particular group at some level may receive a large weight with this procedure. This weighting procedure requires that no cell in the analysis of variance design be empty. For example, given three interfering variables with K, L, and M levels, this procedure requires that

$$n_{jk\ell m} \neq 0$$

for all j, k, ℓ, and m, a more stringent requirement than for Kalton's procedure. Thus, this weighting procedure, although the one usually used, has no advantages over the weighting procedure presented by Kalton (1968); Kalton's procedure does possess several advantages over this alternative procedure.

Indirect standardization -- Indirect standardization has been used even more extensively than direct standardization by demographers, biostatisticians, and other medical researchers, according to Fleiss (1973). The probable reason for the greater usage of indirect standardization is that, unlike the usual form of direct standardization (as presented by Fleiss, 1973), indirect standardization does not suffer from the problem of assigning large weights to unstable cell means and it may be used even if a cell in the design is empty. An example of the use of indirect standardization is again the National Halothane Study, discussed by Moses (1969).

With one interfering variable, indirect standardization is based on the same model as balancing and direct standardization, but employs a different method for estimating parameters. There are two approaches to indirect standardization, both of which have been developed only for one interfering variable. However, for more than one interfering variable, a generalized indirect standardization procedure can be defined in a manner analogous to that for direct standardization, where the design is re-expressed as a design with one interfering variable. The following discussion assumes there to be only one interfering variable, either because this is the original design or because the original multivariable design has been re-expressed.

One approach is that given by Wiley (1973). Whereas balancing obtains estimates by solving the normal equations to arrive at a simultaneous fit to the data, indirect standardization as defined by Wiley obtains estimates in a two-step process. First, estimates of the effect associated with each level of the interfering variable are obtained in a model that assumes no main effect to be associated with group (i.e., the model assumes $\alpha_j=0$, for all j); the model thus is of the form

$$\bar{Y}_{jk} = \mu + \gamma_k + \bar{e}_{jk} \quad .$$

Least squares estimates are obtained for μ and γ_k. These estimates are given by

$$\hat{\mu} = \bar{G} \text{ (the weighted grand mean)} \quad ,$$

and

$$\hat{\gamma}_k = \bar{Y}_k - \bar{G} .$$

Second, these estimates are used to obtain the estimated adjusted score for the j'th group in the model with two main effects,

$$\begin{aligned} \bar{Y}_{jk} &= \mu + \alpha_j + \gamma_k + \bar{e}_{jk} \\ &= \mu_j + \mu_k - \mu + \bar{e}_{jk} \quad . \end{aligned}$$

It now will be shown that when $\hat{\mu}_k$ and $\hat{\mu}$ are constrained to $\bar{Y}_k$ and $\bar{G}$ respectively, the least squares estimate for μ_j is given by

$$\hat{\mu}_j = \bar{G} + \frac{\sum_k n_{jk}(\bar{Y}_{jk} - \bar{Y}_k)}{n_j} \quad ,$$

the estimated adjusted score of Wiley's approach. The squared error for the j'th group is given by:

$$\sum_k n_{jk}(\bar{Y}_{jk} - \hat{Y}_{jk})^2$$

which equals

$$\sum_k n_{jk}(\bar{Y}_{jk} - (\hat{\mu}_j + \hat{\mu}_k - \hat{\mu}))^2 \quad .$$

Substituting $\hat{\mu}_k = \bar{Y}_k$ and $\hat{\mu} = \bar{G}$ and setting the derivative with respect to $\hat{\mu}_j$ equal to zero,

$$-2 \sum_k n_{jk}(\bar{Y}_{jk} - \hat{\mu}_j - \bar{Y}_k + \bar{G}) = 0$$

$$\sum_k n_{jk}(\bar{Y}_{jk} - \bar{Y}_k) + n_j\bar{G} = n_j\hat{\mu}_j$$

$$\bar{G} + \frac{\sum_k n_{jk}(\bar{Y}_{jk} - \bar{Y}_k)}{n_j} = \hat{\mu}_j \ .$$

To compare Wiley's formula with those for balancing and direct standardization, we can compute

$$\hat{\mu}_j - \hat{\mu}_{j'} = \hat{\alpha}_j - \hat{\alpha}_{j'}$$

$$= \left(\bar{G} + \frac{\sum_k n_{jk}(\bar{Y}_{jk} - \bar{Y}_k)}{n_j}\right) - \left(\bar{G} + \frac{\sum_k n_{j'k}(\bar{Y}_{j'k} - \bar{Y}_k)}{n_{j'}}\right)$$

When comparing two groups (the case for which adjustment techniques are most often used), indirect standardization as defined by Wiley provides the following estimate of $\hat{\mu}_1 - \hat{\mu}_2$:

$$\hat{\mu}_1 - \hat{\mu}_2 = \frac{\sum_k [n_2 n_{1k}\bar{Y}_{1k} - n_1 n_{2k}\bar{Y}_{2k} + (n_1 n_{2k} - n_2 n_{1k})\bar{Y}_k]}{n_1 n_2}$$

Since

$$\bar{Y}_k = \frac{n_{1k}\bar{Y}_{1k} + n_{2k}\bar{Y}_{2k}}{n_{1k} + n_{2k}} \ ,$$

$$\hat{\mu}_1 - \hat{\mu}_2 = \sum_k \frac{(n_1 + n_2)n_{1k}n_{2k}(\bar{Y}_{1k} - \bar{Y}_{2k})}{n_1 n_2(n_{1k} + n_{2k})} \ .$$

This can be rewritten as

$$\hat{\mu}_1 - \hat{\mu}_2 = \frac{n_1 + n_2}{n_1 n_2} \sum_k \frac{n_{1k}n_{2k}}{n_{1k} + n_{2k}} (\bar{Y}_{1k} - \bar{Y}_{2k})$$

$$= \left(\frac{n_1+n_2}{n_1 n_2}\right)\left(\sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}}\right)\left(\frac{\sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}}(\bar{Y}_{1k} - \bar{Y}_{2k})}{\sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}}}\right) \ .$$

But the third expression in parentheses is the estimate for balancing. Thus,

$$(\hat{\mu}_1-\hat{\mu}_2)_{\text{Wiley}} = \left(\frac{n_1+n_2}{n_1n_2}\right)\left(\sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}}\right)(\hat{\mu}_1-\hat{\mu}_2)_{\text{Balancing}} \; .$$

Wiley's estimate of the group effect equals the balancing estimate (and thus also equals the estimate from direct standardization) if and only if

$$\frac{n_1n_2}{n_1+n_2} = \sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}} \; ,$$

that is, if and only if

$$\frac{\sum_k n_{1k} \sum_k n_{2k}}{\sum_k (n_{1k}+n_{2k})} = \sum_k \frac{n_{1k}n_{2k}}{n_{1k}+n_{2k}} \; .$$

Another approach to indirect standardization, discussed by Fleiss (1973), involves multiplication of the grand mean estimate by the estimated mean unadjusted score of the j'th group, and division of this product by the mean score that the j'th group would have received if its mean score on the dependent variable within each level of the interfering variable had been the same as that of the population. This approach uses the same estimates as Wiley's approach but combines them in a multiplicative rather than an additive fashion. It yields an estimated adjusted score for the j'th group that can be written as

$$\hat{\mu}_j = \frac{\bar{G}\,\bar{Y}_j}{\dfrac{\sum_k \bar{Y}_k n_{jk}}{n_j}}$$

$$= \frac{\bar{G}\sum_k \dfrac{n_{jk}\bar{Y}_{jk}}{n_j}}{\dfrac{\sum_k n_{jk}\bar{Y}_k}{n_j}} \; .$$

An evaluation of indirect standardization -- Analysis of covariance, balancing, and direct standardization have been compared and contrasted, with a discussion of the relative merits of each. Still to be discussed is the relative value of each of those approaches compared with indirect standardization when adjusting for only one interfering variable.

TABL

Number of Subjects and Mean Score by Group Within Each Level of the Interfering Variable

		Group 1	Group 2	Marginal
Level of Interfering Variable	1	N=1 $\bar{Y}$=20	N=1 $\bar{Y}$=10	N=2 $\bar{Y}$=15
	2	N=1 $\bar{Y}$=40	N=2 $\bar{Y}$=30	N=3 $\bar{Y}$=33.3
	3	N=2 $\bar{Y}$=60	N=3 $\bar{Y}$=50	N=5 $\bar{Y}$=54
	4	N=9 $\bar{Y}$=80	N=4 $\bar{Y}$=70	N=13 $\bar{Y}$=77
	5	N=2 $\bar{Y}$=100	N=5 $\bar{Y}$=90	N=7 $\bar{Y}$=92.9
Marginal		N=15 $\bar{Y}$=73.3	N=15 $\bar{Y}$=63.3	N=30 $\bar{Y}$=68.3

Two examples are cited to illustrate certain features of the two alternative estimation procedures for indirect standardization. First, consider the fictitious data given in Table I. Within each group, there is a perfect linear relationship between the interfering variable and the dependent variable. Also, at each level of the interfering variable, Group 1 has a mean score 10 points higher on the dependent variable than does Group 2. Note also that the frequency distribution of scores on the interfering variable is quite different for the two groups.

TABLE II

Adjusted Mean Scores from Table V

	Group 1	Group 2	Difference
Wiley	72.71	63.96	8.75
Fleiss	72.67	63.92	8.75
ANCOVA	73.33	63.33	10
Balancing	73.33	63.33	10

Adjusted mean scores for the two groups as derived from each adjustment technique are given in Table II. Analysis of covariance and balancing both estimate the difference between adjusted scores to be 10, while either method of indirect standardization estimates the difference to be slightly less than 10. But the evidence is that, within any level of the interfering variable, the difference is in fact 10, so this is the desired difference between the adjusted scores. Analysis of covariance and balancing both recover this difference, in contrast to both forms of indirect standardization.

As a general rule, analysis of covariance will recover the desired difference whenever there is no interaction between the interfering variable and the grouping and all higher-order trends are zero in the data. Balancing will recover this desired difference whenever there is no interaction between the interfering variable and the grouping. When there is such an interaction, the mean score difference between groups varies depending on the particular level of the interfering variable, so there is varying evidence on the difference between the groups. In general, indirect standardization will not recover this desired difference.

A second example shows another way in which either approach to indirect standardization may yield misleading results. Consider the data in Table III. For these data, Group 1 has a mean score of 60 at every level of the interfering variable, while Group 2 has a mean score of 40 at every level. In addition, there is little overlap between the two groups on the interfering variable; only for level 3 are there observations for both groups, and here, as elsewhere, Group 1 has an average score of 60 while Group 2 has a mean score of 40. It seems reasonable to conclude that within groups the interfering variable and the dependent variable are unrelated; instead, group 1 members tend to score higher than Group 2 members on both the interfering variable and the dependent variable.

TABLE III

Number of Subjects and Mean Score by Group Within Each Level of the Interfering Variable

		Group 1	Group 2	Marginal
Level of Interfering Variable	1	N=0	N=50 $\bar{Y}$=40	N=50 $\bar{Y}$=40
	2	N=0	N=25 $\bar{Y}$=40	N=25 $\bar{Y}$=40
	3	N=25 $\bar{Y}$=60	N=25 $\bar{Y}$=40	N=50 $\bar{Y}$=50
	4	N=25 $\bar{Y}$=60	N=0	N=25 $\bar{Y}$=60
	5	N=50 $\bar{Y}$=60	N=0	N=50 $\bar{Y}$=60
Marginal		N=100 $\bar{Y}$=60	N=100 $\bar{Y}$=40	N=200 $\bar{Y}$=50

Both analysis of covariance and balancing support this conclusion, as seen in Table IV. The adjusted score for each group equals the unadjusted score for the group. Indirect standardization, however, gives the impression that the difference between the mean scores of the groups can be explained by their being different on the interfering variable. The adjusted rates obtained from indirect standardization are very nearly equal for the two groups.

TABLE IV

Adjusted Mean Scores from Table III

	Group 1	Group 2	Difference
Wiley	52.50	47.50	5.00
Fleiss	52.17	47.06	5.11
ANCOVA	60	40	20
Balancing	60	40	20

Adjusted mean scores obtained from balancing will differ from unadjusted mean scores if and only if the interfering variable is related to the dependent variable within each group (homogeneity of the relation is assumed) and the groups have different means on the interfering variable. For analysis of covariance, there is the additional restriction that adjusted scores will differ from unadjusted scores only to the extent of linear relationship. Indirect standardization, on the other hand, may give adjusted scores that are different from the unadjusted scores despite groups having the same mean on the interfering variable (see Tables I and II), or despite there being no (within group) relation between the interfering variable and the dependent variable (see Tables III and IV).

With reference to admission data, let us consider an example of how either approach to indirect standardization may be misleading. Suppose that in Table III, Group 1 represents male applicants, Group 2 represents female applicants, and the interfering variable is height. In these hypothetical data, male applicants all have heights which place them in level 3 or 4 or 5, while female applicants all are placed in level 1 or 2 or 3. Male applicants have an average admission rate of 60 percent regardless of their height, while female applicants have an average of 40 percent regardless of their height. There is no relation between height and admission within sex, i.e., an admission committee does not act on the basis of an applicant's height. Thus, in all probability, if the average height of women were to increase, their admission rate would stay the same. But indirect standardization leads us to believe that if women were only taller, they would be accepted at almost the same rate as men. From the results of indirect standardization, it would seem that males and females are being accepted at nearly the same rate, once we take into account the difference in average height. But results obtained from balancing and analysis of covariance will yield an adjusted admission rate for male applicants of 60 percent and a rate for females of 40 percent, implying that the admission rate for males would remain substantially higher than the rate for females even if the average height for female applicants were to increase.

The example given in Table III is an extreme case illustrating a possible difference between results obtained by indirect standardization and results from balancing and analysis of covariance. Both of the latter techniques rely on the relation between the interfering variable and the dependent variable <u>within each group</u>. Such a relation hould be found (except for chance error) if and only if the interfering variable is exerting an influence on the decisions of the admission committee, in which case it is reasonable to predict that if a group's mean score on the interfering variable were higher, the group's admission rate also would be higher. When it is desired that adjustment be made only for such a "within-group" relation, the use of either balancing or analysis of covariance is always preferable to the use of indirect standardization.

In general, indirect standardization seems to offer no advantages over balancing, but seems to suffer from several disadvantages. The only advantage indirect standardization has over analysis of covariance is that it allows for a nonlinear relation between the interfering and the dependent variable, but balancing also makes this allowance. Thus, it seems that either balancing or analysis of covariance should be used to obtain adjusted scores.

Chapter V: "Smear and Sweep" Analysis

One of the secondary objectives of this grant was the investigation of other data analytic techniques used to adjust for nuisance confounding in the NAEP studies. One such technique (and the only other "non-standard" technique of major consequence) is that known as smear-and-sweep. The following chapt gives the basic results on smear-and-sweep and its relation to balancing and the nonorthogonal analysis of variance.

In many behavioral or social research situations, researchers may want to estimate treatment effects or the relationships between input and output variables, while controlling for a number of extraneous variables. Some methods of analysis use the input and the extraneous variables to form multifactor classifications, and then estimate the treatment effects or relationships, adjusted for the effects of extraneous variables. A large number of variables available in the data may thus be selected to form multifactor cross-classifications, resulting in few observations per cell, indeed some cells in the crossed-classified may have no observations. For example, if 2500 sixth-grade students are involved in a study of educational progress, and later stratified into subgroups by region of the country (four levels), sex, race (three levels), type of community (seven levels), and parental education (five levels), they will be distributed over 840 cell combinations giving an average of about three observations per cell. With this many cells the data in each cell become too sparse to allow stable estimates of cell values, and direct control on all the extraneous variables by multifactor classification may, therefore, be impractical.

Smear and Sweep Analysis

One method developed for the d[illegible]ed data resulting from the after-the-fact classifications is the Smear-and Sweep analysis. This method first appeared in the report of the National Halothane Study (Gentleman, Gilbert & Tukey, 1969) in which the death rate of the patients in operations using Halothane was examined. It was later considered by the National Assessment of Educational Progress (NAEP) as a possible method to obtain sharper subpopulation weights (see Ahmann, 1973, pp. 108-109).

Smear-and-sweep is a method that first pools the cells of the control variables in a cross-classification table into categories in which the cell values (e.g., proportions, ratios, means) are reasonably similar, and then calculates and compares the "effects" for an independent variable of interest across a final set of categories. The basic strategy is to form a two-way table on two of the control variables at a time. (This step is referred to as smearing.) The cells of this table are then ordered on a single dimension according to the value of the dependent variable in the cells. The value of this dependent variable may be simply the observed data, Least Squares estimated, or statistically adjusted values. Then the adjacent cells are pooled into a smaller number of categories which define the levels of a new conglomerate variable. (This step is referred to as sweeping.) The process is then repeated by forming another classification table consisting of the newly formed conglomerate variable and another control variable. This process is continued until only a single conglomerate variable remains. The final conglomerate variable is then cross-classified with the independent variable of interest. This table is then used to compute marginal estimates and perform some comparisons among the levels of the independent variable of interest, using classical techniques such as analysis of variance.

The essence of this method is that it permits the researcher to handle many known and available variables as control variables. The process of Smear-and-Sweep will presumably control or minimize the effects of extraneous variables, and thus allows better estimates of the effects due to the independent variable of interest. By using two variables at a time, the number of observations in each cell combination may be large enough for stable estimates.

An Illustration

Smear-and-Sweep analysis is illustrated by the following hypothetical data set. A probability sample of 1,933 high school graduates were given a science test. Their test scores were scored either 1 (pass) or 0 (fail).

Suppose that, using this data set, a researcher was interested in testing ethnic group on the students' test scores after contro ng for sex, region, socioeconomic (SES), and high school curricular program (HSP). The researcher could use a multifactor design, and apply analysis of variance to obtain adjusted marginal estimates for ethnic groups; however, in so doing, the observations within each cell combination would be very sparse. Many cells would have two or three observations while some other cells would have none; the cell sizes might not be sufficient to provide stable estimates.

Given these problems, the researcher chose to use a more "data analytic" approach, namely Smear-and-Sweep. The researcher first cross-classified students on the basis of their socioeconomic background and high school program (HSP). SES had three categories: high, middle, and low in correspondence with upper quartile, middle two quartiles, and lower quartile of the SES composite scores, respectively. HSP was defined by college preparatory (academic), general and vocational-technical (voc-tech) programs. The proportion of pass for each cell combination was computed as follows:

$$P_{ij} = \frac{S_{ij}}{N_{ij}}$$

where S_{ij} is the number of students who had a score of 1 in the i'th SES and j'th HSP, and N_{ij} is the total number of respondents in this cell.

The obtained proportions were then ordered, and their corresponding cells were grouped into five categories as indicated in Table 1. The criterion for grouping was that the range of proportions in each category should not exceed .05. These five categories comprised a new "conglomerate" variable; each category incorporating some "effects" due to SES and HSP.

Table 1

Proportion of Pass for SES and HSP
Cross-Classification Groups

SES	High School Program	Proportion of Pass	Category
High	Academic	.82	1
Middle	Academic	.67	2
High	General	.56	3
Low	Academic	.55	3
High	Voc-Tech	.34	4
Middle	General	.32	4
Middle	Voc-Tech	.22	5
Low	General	.22	5
Low	Voc-Tech	.17	5

The newly formed conglomerate variable was then cross-classified with four geographic regions and resulted in a four-by-five classification table. The cell values (i.e., proportions) of this table were calculated, and the cells were grouped in accordance with the rules described previously. The results are presented in Table 2. As seen in the table, the new conglomerate variable included seven categories as indicated by the number in the parentheses.

Table 2

Proportion of Pass for Region and the
First Conglomerate Variable Combination Group

	Conglomerate Variable				
Region	1	2	3	4	5
Northeast	.84(1)*	.70(2)	.60(3)	.26(7)	.26(8)
North Central	.82(1)	.65(2)	.52(5)	.30(7)	.21(8)
South	.82(1)	.64(3)	.56(4)	.36(6)	.19(8)
West	.80(1)	.67(2)	.58(4)	.34(6)	.30(7)

*The figures in parentheses denote the levels of newly formed variables.

Similarly, the second newly formed variable was then crossed with two sex groups to for a two-by eight classification table. The proportion of pass in each cell combination is presented in Table 3. Again, the proportions were ordered, and their corresponding cells were grouped into categories, as indicated by the number in the parentheses, based upon the same criterion that the range of proportions in each category should not exceed .05.

Table 3

Proportion of Pass for Sex and the Second Conglomerate Variable Combination Group

Sex	1	2	3	4	5	6	7	8
Male	.82(1)*	.69(2)	.62(3)	.53(4)	.53(4)	.34(5)	.30(6)	.20(7)
Female	.83(1)	.67(2)	.64(2)	.60(3)	.50(4)	.36(5)	.28(6)	.18(7)

*The figures in parentheses denote the levels of newly formed variable.

The last newly formed conglomerate variable was then cross-classified with ethnic group, which was the independent variable of interest. There were four ethnic groups: black, white, Hispanic (Spanish American), and others. The resulting four-by-seven table and its cell values are presented in Table 4. The last column of the table presents the adjusted average of cell proportions for each ethnic group. No substantial differences among ethnic groups were revealed, although whites had a slightly lower proportion than other groups. It should be noted, however, that these adjusted estimates were quite different from unadjusted ones. Had the proportions been estimated without controlling for sex, region, SES and HSP, the estimates would have been .36, .47, .38, and .42 for blacks, whites, Hispanics and others, respectively. Whites would have had a much higher proportion of pass than blacks.

Table 4

Proportion of Pass for Race and the
Third Conglomerate Variable Combination Group

Ethnic				Conglomerate Variable				Adjusted
Group	1	2	3	4	5	6	7	Average
Black	.87	.66	.59	.38	.44	.43	.23	.52
White	.82	.67	.61	.55	.34	.28	.17	.49
Hispanic	.79	.65	.60	.57	.40	.29	.24	.51
Other	.89	.74	.65	.51	.39	.25	.18	.52

It is seen that the entire process of smear-and-sweep requires the selection of classification variables, and the following guidance functions: (1) the order in which the classification variables to be presented in the analysis, and (2) the criterion for cell pooling. The pooling criterion may be that each category contains (1) an approximately equal number of pass or fail, (2) an equal number of sample members, (3) equal variance of estimated cell values (Gentleman, Gilbert, & Tukey, 1969, p. 289), or (4) equal range of cell values. Once the guidance functions are sufficiently determined, the computational procedures become straightforward.

It should be noted that the cell values in the previous cross-classification tables were estimated simply by using the observed data. Other estimating procedures are possible. For example, one might use the formula

$$P_{ij} = \frac{\Sigma_n W_{ijn} x_{ijn}}{\Sigma_n W_{ijn}} .$$

where W_{ijn} is the sample weight for the n'th individual in the ij'th cell, and x_{ijn} is the individual's score, either 1 or 0, 1 being pass, 0 being fail.

Some Considerations to Smear-and-Sweep Analysis

Although smear-and-sweep has been applied to the analysis of the National Halothane Study (Gentleman, Gilbert, & Tukey, 1969), no proof of the stability and accuracy of estimation has been given. Many questions involving the choice of guidance functions such as the number of categories, and the order of the classification variables introduced into the process, are unanswered. Among such questions, the following ones are considered critical:

1. Does the number of categories selected affect the stability and accuracy of the estimates?
2. Does the order of treating the interfering variables affect the estimation of the effects of the independent variable?
3. How do the results obtained by smear-and-sweep differ from those obtained by classical ANOVA?

To answer these questions, three sets of hypothetical data were constructed. Each set of data was derived by using the following four-factor main-effect model:

$$Y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + \theta_\ell + \varepsilon_{ijk\ell}$$

in which $\mu = \quad , \Sigma\alpha_i = \Sigma\beta_j = \Sigma\gamma_k = \Sigma\theta_\ell = 0$, and $\varepsilon_{ijk\ell} \sim n(0,1)$. This additive model was selected for its simplicity. If smear-and-sweep does not work in such a simple model, it will very likely fail in a more complicated non-additive model.

In the four-factor model, the first factor (independent variable), denoted by A, is the variable of interest. A has two levels; thus, the estimates of effects for A_1 and A_2 are of main concern. The other three factors, designated by B, C, and D, respectively, are referred to as interfering variables. All these variables are assumed to be associated with the dependent variable.

The error components for observations in each cell were chosen to be normally distributed with a mean of 0 and a standard deviation of 1, and were generated using a standard method (Box & Muller, 1958). The main effects for all factors in the analysis were fixed at the values presented in Table 5. These values represent differences ranging from four to one-tenth standard deviations apart.

Table 5

Main Effects Selected for Each Set of Data

Level		Data Set I	 II	 III
A	1	2.00	1.50	1.00
	2	-2.00	-1.50	-1.00
B	1	-1.50	-1.00	- .50
	2	1.50	1.00	- .50
C	1	1.00	.50	.10
	2	-1.00	- .50	- .10
D	1	.50	.10	.05
	2	- .50	- .10	- .05

The cell frequencies (i.e., number of observations in each of the cell combinations) are not equal, reflecting situations likely to be confronted in actual studies. These frequencies, as presented in Table 6, were arbitrarily chosen, with only the restriction that there be sufficient degrees of freedom for testing any main effect.

A. Number of Categories

In the sweeping process, a critical question is: How many categories should one use? It has been suggested that a relatively large number of categories would be preferred (Gentleman, Gilbert & Tukey, 1968, p. 296). However, results in the National Halothane Study and the National Assessment of Educational Progress did not show a significant difference resulting from the number

Table 6

Cell Frequencies

A	B	C	D		Cell Frequency
1	1	1	1		5
1	1	1	2		4
1	1	2	1		3
1	1	2	2		2
1	2	1	1		2
1	2	1	2		3
1	2	2	1		4
1	2	2	2		5
2	1	1	1		3
2	1	1	2		2
2	1	2	1		4
2	1	2	2		5
2	2	1	1		4
2	2	1	2		3
2	2	2	1		5
2	2	2	2		2
				Total	56

of categories. This may be due to the fact that the cell values in those studies were so homogeneous that different grouping processes would not be sensitive enough to affect estimates for each category. Nevertheless, differential effects resulting from various numbers of categories were investigated with the following procedures.

First, factors B and C were smeared and swept into a new variable with four categories. This new variable was then smeared over factor D and resulted in a two by four table. The estimated cell values were ordered and are presented in Table 7.

Table 7

Estimated Cell Value

	Cell Order							
Data Set	1	2	3	4	5	6	7	8
I	2.374	1.932	.505	.063	-.024	-.466	-1.894	-2.336
II	1.332	.974	.463	.105	-.066	-.424	- .936	-1.294
IV	.482	.413	.084	.024	.014	-.045	- .374	- .444

The ordered eight cells were swept into categories, starting from the cell with the highest value, in accordance with each of the following criteria:

1. Cells with positive values would be swept into one category, whereas those with negative values would be swept into another.
2. The range of cell values within each category would be less than .45.
3. The range of cell values within each category would be less than .30.
4. The range of cell values within each category would be less than .05.

The numbers of resulting categories formed for each data set are presented in Table 8.

Table 8

Number of Categories Formed Under Four Criteria

	Criterion			
Data Set	1	2	3	4
I	2	4	7	8
II	2	4	7	8
III	2	3*	3	7

*This classification was not used in the subsequent analyses

The independent variable A was then cross-classified with each final conglomerate variable to form a two-way classification table. Analysis of variance was then conducted for this two-way classification table, and the adjusted marginal means for A_1 and A_2 were computed with an additive model. The difference between the two levels, as contrasted with those expected true differences (see Table 5) and those estimated by multifactor ANOVA, are presented in Table 9. Some smear-and-sweep estimates (e.g., those obtained by two categories) are as close to the expected differences as those obtained by multifactor ANOVA. The number of categories does affect estimates of the differences. For data set I, the more categories used, the smaller the difference between A_1 and A_2, and the greater the deviation of the estimated difference from the expected value. This finding contradicts the suggestion that a larger number of categories be used (Gentleman, Gilbert & Tukey, 1969, p. 296). However, this finding of the number of categories being negatively related to the magnitudes of the estimates is not necessarily supported by results from data set III, for which the seven-category estimate is closer to the expected than the three-category estimate. It is, therefore, not clear how systematically the choice of the number of categories can affect the precision of estimation. The authors suspect that the effects may fluctuate randomly. When the right number of categories is "hit," the estimates obtained by smear-and-sweep analysis can be as good as those by ANOVA or other methods.

B. Order of Variables

It has been argued that the order of the presentation of the variables might be analogous to the step-wise regression analysis in which the most important variable should be introduced first (Gentleman, Gilbert, & Tukey, 1969, p. 295). Previously, however, no systematic examination of this argument has been conducted. It is, therefore, the purpose of this portion of the study to explore the order effect of variables in the smear-and-sweep process.

Table 9

Difference Between A_1 and A_2
As Obtained from Various Analyses

Analysis		Data Set	
	I	II	III
Smear and Sweep			
AxN(2)*	4.039	3.025	1.965
AxN(3)			1.433
AxN(4)	3.799	2.871	
AxN(7)	3.764	2.747	1.760
AxN(8)	3.762	2.762	
Factorial ANOVA	3.938	2.938	1.938
Expected	4.000	3.000	2.000

*The number in the parenthesis indicates the number of categories for the final conglomerate variable.

For the same design and data used in the previous section, three possible orders of variable presentation were investigated. They are:

(1) B, C, D (the same as C, B, D),

(2) D, B, C (the same as B, D, C), and

(3) C, D, B (the same as D, C, B).

The alphabetic order of B, C, and D indicates the order of importance of these variables in terms of the magnitude of their effects (see Table 5).

Estimated differences between A_1 and A_2 from data set I under two cell-pooling criteria are presented in Table 10. The results do not support the argument that the most important variables should be introduced first. Results from the other two sets of data also failed to provide positive evidence. It seems that what makes estimates different is not the order of variable presentation but the resulting number of final categories.

Table 10

Estimated Difference Between A_1 and A_2
With Three Orders of Variable Presentation for Data Set I

Order of Presentation	Cell-Pooling Criterion	
	Range less than .05	Range less than .30
B, C → D	3.762(8)*	3.745(4)
B, D → C	3.762(8)	3.764(7)
C, D → B	3.760(6)	3.799(4)

*The figures in parentheses indicate the number of categories of the final conglomerate variable.

C. Comparisons on Test Statistics

Analysis of variance may be applied to the final cross-classification table to test the significance of treatment effects or group difference (Gentleman, Gilbert, & Tukey, 1969). The question then is: To what extent will the results obtained by smear-and-sweep differ from those obtained by a factorial analysis of variance if the data permit the latter analysis? To answer this question, nonorthogonal analysis of variance (ANOVA) for a factorial design was performed on data used in previous sections to obtain test statistics for A eliminating B, C, and D (A|B, C, D); namely, unconfounded test of A (see Appelbaum & Cramer, 1973). Nonorthogonal ANOVA was also conducted on the final two-way cross-classification table resulting from the smear-and-sweep process, with A as one dimension and the newly formed variable as another dimension. (It should be noted that the order of control variables introduced into the smear-and-sweep process was B, C, then D.)

The mean squares and degrees of freedom for each test are presented in Table 11. It can be seen that in the smear-and-sweep analyses, between-group variance decreases, as expected, as the number of categories of the final conglomerate variable increases. The within-group variances, however, fluctuate. Converting the variances into F statistics, all of them are significant at the .01 level with their associated degrees of freedom. As far as significance testing is concerned, smear-and-sweep provides results similar to factorial analysis of variance. However, smear-and-sweep analysis may provide a more conservative test. Comparing A|B, C, D, and A|N(8), for example, both designs produce the same magnitude of error variance, same degrees of freedom for A effects, but their between-group variances are quite different; A|N(8) has much smaller between-group variance than A|B, C, D. It is possible when A effects are small, that the A|N(8) may provide test statistics indicating non-significant A effects while A|B, C, D indicates significant differences.

Summary and Discussion

Smear-and-sweep analysis is a method to compute summary statistics such that the effects of interfering variables are reduced or controlled. The basic strategy is to pool cells of similar values into categories. It involves the following steps: (1) forming a two-way classification table (i.e., smearing) and estimating cell values; (2) forming categories based on cell values (i.e., sweeping); and (3) comparing the values among levels on the interested independent variable across the final set of categories.

Since its development and application in the National Halothane Study, it has received little systematic evaluation. This study shows that the precision of the summary statistics depends very much on the choice of the number of categories; however, it seems that it is not always preferable to have a large number of categories. The investigation does not support the argument that the greater the number of categories, the better the estimates. Furthermore, the choice of the categories has not yet been systematically defined. Cluster analysis could be an alternative to sequential two-way aggregation. Further investigation is warranted.

Table 11

Comparisons on Test Statistics

Design	Source	Mean Squares			d.f
		Data I	Data II	Data III	
Factorial					
ANOVA	Between	214.85	119.59	52.03	d.f. = 1,40
A|B, C, D	Within	1.55	1.55	1.55	
Smear-and Sweep*					
	Between	228.49	128.16	54.64	d.f. = 1,52
A|N(2)	Within	2.41	1.71	1.56	
A|N(4)	Between	191.70	109.53		d.f. = 1,48
	Within	1.51	1.48		
A|N(7)	Between	182.30	98.24	39.79	d.f. = 1,42
	Within	1.50	1.49	1.49	
A|N(8)	Between	116.84	17.91		d.f. = 1,40
	Within	1.55	1.55		
A|N(3)	Between			45.76	d.f. = 1,50
	Within			1.39	

[*] The number in the parentheses indicates the number of categories for the final conglomerate variable.

The suggestion of introducing the most important interfering variable first in the process is also not supported. The order itself does not seem to be a determinative factor in the precision of estimates. It is the number of resulting categories that affects estimates. Once the number of categories of the final conglomerate variable is selected, the order of variable presentation does not seem to be critical. However, it should be noted that the order of presentation may very likely determine the selection of the number of categories.

The results of the investigation also show that smear-and-sweep tends to provide a conservative significance test as compared to factorial analysis of variance. When data are sparse, smear-and-sweep is an alternative method that may lend some strength to stable estimates, and explore the treatment effect or possible relationships between classification and dependent variables.

Chapter VI: A Comparison of Balancing and Analysis of Covariance in the Adjustment of Educational Data.

Female and male admission [illegible] graduate programs at the University of North Carolina [illegible] are compared for 1972-73 and 1973-74. To assess possible sex-related bias in admission, rates are adjusted for applicant qualifications by analysis of covariance and by balancing.

The adjusted admission rates reflect, in one case, i.e., for one program and one admission year, a slight advantage for male applicants over females, while in three cases, female applicants were granted a slight advantage over males in admission. In the remaining four cases, there is no evidence that sex of applicant, per se, played a role in admission decisions. Wherever a sex-related advantage is detected, the favored sex is that with the fewer applicants to the program.

The dependent variable in this study, defined for each applicant for a given program and enrollment year is

$$Y_{ij} = \begin{cases} 1, \text{ if admitted} \\ 0, \text{ if rejected} \end{cases} ,$$

where $j=1$ if the applicant is female and $j=2$ if the applicant is male, and $i=1, 2, \ldots, n_j$ is the number of female (or male) applicants to the program for that year. Then

$$P_j = \sum_{i=1}^{n_j} (Y_{ij}/n_j)$$

is the mean Y_{ij} for sex j, and also represents the proportion of applicants of sex j who were admitted. The P_j values are the female and male admission rates presented in Table I.

Given in Table I are the unadjusted rates of admission by sex, and in Table II the within-group correlations of interfering variables with admission

TABLE I

Graduate Admission Rates by Sex for 1972 and 1973*

Field	Year	Female Admission Rate	Number of Female Applicants	Male Admission Rate	Number of Male Applicants
English	1972	32.1	165	34.0	235
	1973	20.3	153	25.0	204
History	1972	68.2	44	55.5	182
	1973	54.5	55	48.0	177
Library	[illegible]	[illegible]4	175	70.2	57
Science	973	[illegible]1.0	157	50.0	40
Sociology	1972	31.6	38	18.2	66
	1973	22.6	31	11.6	69

*Excluded from the table are all applicants for whom less than complete data were available from the set of undergraduate grade point average, GRE scores, and two letters of recommendation.

TABLE II

Point-Biserial Correlations of Qualification Variables with Admission for Female (F) and Male (M) Applicants

Field	Year	GPA		GRE V		GRE Q		GRE Adv		REC	
		F	M	F	M	F	M	F	M	F	M
English	72	.29	.42	.16	.46	.18	.34	.20	.33	.26	.22
	73	.37	.27	.28	.29	.15	.27	.25	.35	.17	.19
History	72	.52	.52	.56	.43	.24	.43	.39	.33	.42	.49
	73	.38	.36	.62	.54	.43	.47	.14	.41	.42	.35
Library	72	.35	.42	.50	-.06	.30	-.07	--	--	.39	.29
Science	73	.43	.38	.38	.55	.48	.65	--	--	.33	.47
Sociology	72	.27	.46	.51	.14	.67	.15	.60	.16	.07	.32
	73	.48	.12	.16	.34	.06	.35	.37	.26	.39	.15

Also of interest are the mean differences between the sexes on the interfering variables, given in Table III. These are computed by subtracting the mean for female applicants from the mean for male applicants, so that a positive mean difference represents a male advantage and a negative mean difference represents a female advantage. The unit, in each case, is that in which each variate is naturally recorded.

Somewhat more informative are the standardized mean differences, presented in Table IV. Here, each male-female mean difference is divided by the standard error of the mean difference. Each value in Table IV represents a _t_ statistic. Those values which differ from zero by approximately two or more are judged [illegible] zero sufficiently to represent a statistically significant difference between sexes. The results of Table IV represent the values of

$$t_m = \frac{\bar{x}_2^{(m)} - \bar{x}_1^{(m)}}{\sqrt{\frac{s_m^2}{n_2} + \frac{s_m^2}{n_1}}}$$

The index m identifies the covariate, as defined in Table V; s_m is the within-sex standard deviation for that covariate; n_j indicates the number of applicants of [illegible] j.

Inspection of Tables III and IV is instructive. Without exception, the mean grade point average for women applicants is higher than that for males for each program and each year. On GRE scores, women applicants show higher mean scores than male applicants on the verbal test (except for applicants to the Department of English), while males show higher means than females on the quantitative test, and also (with the exception of Sociology applicants in 1973) display higher mean scores on the advanced test. For each program and each year, the mean summary score derived from letters of recommendation is higher for males than for females. The mean differences on GRE-Q for male and female applicants to English is extraordinarily large, more than 70 points in both years (Table III), with highly significant _t_ statistics, 7.1 and 6.6 (Table IV).

A comment is in order concerning the consistent advantage of male applicants on mean level of recommendation for graduate study (Tables III and IV), especially since it contrasts with a female advantage on grade point average and (usually) on GRE-V.

TABLE III

Male-Female Mean Differences for Applicants on the Interfering Variables

Interfering Variable	English 1972	English 1973	History 1972	History 1973	Library Science 1972	Library Science 1973	Sociology 1972	Sociology 1973
GPA	- .108	- .040	- .035	- .192	- .206	- .111	- .149	- .061
GRE-V	8.27	17.36	-42.71	-30.22	- 2.18	- .32	-25.40	-25.58
GRE-Q	70.68	71.04	15.17	41.77	27.89	16.38	24.01	13.56
GRE-A	7.39	19.96	39.90	36.32	--	--	2.39	-11.81
REC.	.041	.063	.071	.137	.306	.087	.133	.164

TABLE IV

Standardized Male-Female Mean Differences on the Interfering Variables

Interfering Variable	English 1972	English 1973	History 1972	History 1973	Library Science 1972	Library Science 1973	Sociology 1972	Sociology 1973
GPA	-2.77	- .95	- .49	-3.00	-2.75	-1.59	-1.64	- .73
GRE-V	.91	1.75	-2.65	-1.94	- .13	- .02	-1.41	-1.26
GRE-Q	7.12	6.64	.80	2.47	1.72	.84	1.01	.55
GRE-A	.93	2.34	3.16	3.22	--	--	.12	- .55
REC.	.84	1.24	.71	1.88	3.48	.93	1.31	2.28

TABLE V

Variables Pertaining to Admission Qualifications

Variable	The Nature of the Variable
X_1	Undergraduate grade point average for final two years (GPA)
X_2	Verbal score on the Graduate Record Examination (GRE-V)
X_3	Quantitative score on the Graduate Record Examination (GRE-Q)
X_4	Score on the Advanced Test, Graduate Record Examination (GRE-A)
X_5	Mean recommendation (with each coded 0-4)

It is possible that this represents a bias toward males on the part of those who recommend applicants, who most frequently are male faculty members. However, the recommendation is couched in terms of the probability that the candidate will successfully complete a doctoral program; the apparent male advantage could be the result of possibly valid judgments that women have been more likely than men to discontinue graduate work before receiving the Ph.D.

TABLE VI

Comparison of Unadjusted and Adjusted Female (F) and Male-Female (M-F) Admission Rates

Field	Year	Unadjusted		Adjusted			
				ANCOVA		Balancing	
		F	M-F	F	M-F	F	M-F
English	1972	32.1	1.9	33.5	- .4	31.9	2.3
	1973	20.3	4.7	23.0	- .1	22.6	.7
History	1972	68.2	-12.7	67.9	-12.3	68.3	-12.8
	1973	54.5	- 6.5	48.9	.9	46.3	4.3
Library Science	1972	59.4	10.6	59.6	10.1	59.0	12.4
	1973	51.0	- 1.0	50.6	.8	50.0	3.8
Sociology	1972	31.6	-13.4	29.2	- 9.7	30.4	-11.5
	1973	22.6	-11.0	21.9	-10.0	22.3	-10.6

A reasonable indication of apparent favoritism toward males in admission to graduate study is provided by the male-female difference in admission rate. For each program and each year, this difference is compared in Table VI with the male-female difference after adjustment by analysis of covariance and adjustment by balancing for the sex differences on all five interfering variables. The adjusted differences in admission rate using analysis of covariance are plotted against the unadjusted differences in Figure 1.

From Table VI or Figure 1, several conclusions follow.

First, neither the covariance adjustment nor the balancing adjustment radically changes the impressions gained from assessing unadjusted differences in admission rates for men and women. In the most extreme cases, History in 1972 and Sociology in both years appeared to favor female applicants, and the appearance applies to admission rates following adjustment; Library Science in 1972 appeared to favor male applicants and, again, the adjustment leads to no different appearance.

A somewhat different conclusion does arise, however, regarding the influence of sex of applicant upon admission policy in History for 1973 and in English for 1973. The adjusted results for History, 1973, suggest that the apparent favoritism of female applicants may have been a cor[illegible] male-female differences in scores on the covariates. While the unadjusted rates favored women by 6.5 percent, adjusted rates actually favor men by .9 percent from ANCOVA, and by 4.3 percent from balancing. For English, 1973, unadjusted rates suggested a tendency to favor males slightly over females in admission. After adjustment, there are only negligible differences between male and female rates.

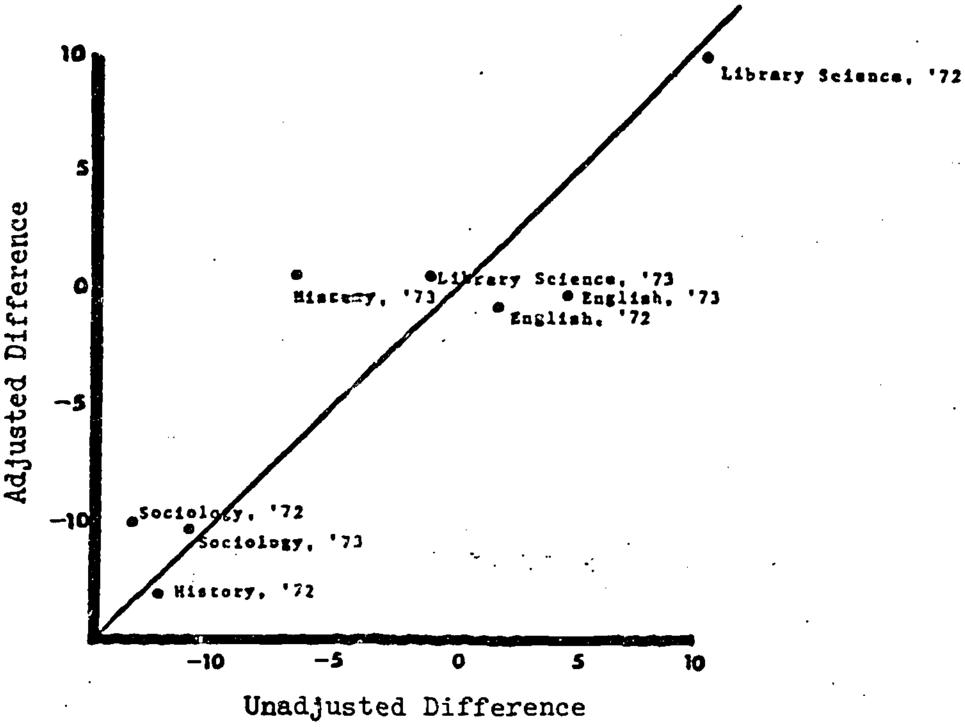

FIG. 1

Male-Female Differences in Admission Rates: Adjusted Differences (by Analysis of Covariance) vs. Unadjusted Differences

The fourth and sixth columns of Table VI suggest that, for one program in one year, Library Science in 1972, male applicants were accepted more frequently than female applicants for reasons other than differences in grade point average, test scores, or strength of recommendations. For History in 1972 and for Sociology in both years, female applicants appear to have been granted a similar advantage. In the remaining four cases, there is no evidence that sex of applicant, per se, played a role in admission decisions.

DISCUSSION

Adjustment techniques as used in this study provide an answer to the question of what female and male admission rates to graduate study might have been if the two sexes had presented equal qualifications on a set of interfering characteristics, as these characteristics were used by the admission committee to select applicants. Interpretation of adjusted scores must be tempered by the realization that these adjustments occur under the admission committee's definition of qualification.

Adjusted rates provide some clue as to what male and female acceptance rates might have been if males and females had had the same distributions for the interfering variables. It is most meaningful to examine adjusted rates in conjunction with unadjusted rates, which represent how often females and males in reality were accepted. Adjusted acceptance rates provide more information concerning the fairness of the admissions committee, but when examined in conjunction with unadjusted rates they also provide information concerning the differential qualifications of males and females. A large difference between unadjusted rates and rates adjusted for a particular characteristic suggests that the average score is quite different for male and female applicants, and also that the committee considers this difference to be important. Such a pattern of scores may provoke interest as to why the applicants of one sex are more qualified than those of the other on the average and also as to why the committee considers this characteristic to be important in defining qualification. For example, females applying to these four departments seem to have higher GPA and GRE-V averages but lower GRE-Q, GRE-ADV, and REC. averages than male applicants in the corresponding departments. It would be interesting to investigate how uniform this pattern is among applicants to other departments at this university and among applicants to other graduate schools. Even if adjusted male and female rates are approximately the same, further investigation may be desirable if unadjusted rates are very different from the other, since this suggests that the applicants of one sex are much more qualified than those of the other. For example, a validity study might be conducted to insure that interfering characteristics are being used fairly, or a study might be done to determine why highly qualified persons of one sex but not of the other are motivated to apply to the department.

In this investigation of admission of applicants to four graduate programs, only modest differences were observed between admission rates for females and males, sometimes favoring one sex and sometimes favoring the other. After adjusting for sex differences in undergraduate grade point average, Graduate Record Examinations scores, and recommendations, some of these differences remained (three favoring females, one favoring males), while others disappeared. The study illustrates the appropriateness of adjusting admission rates before drawing conclusions concerning sex differences in admission to graduate study.

Conclusions

A recurring problem in educational research (and indeed in social, behavioral, and medical research) has been the adjustment of data to account for initial differences among observed groups of individuals on attributes uncontrollable by the researcher. Unlike the experimental solution introduced by Fisher--randomization--which typically cannot be employed in the educational setting, the majority of "solutions" employed by educational evaluators have been essentially statistical or data analytic adjustments. While the use of this class of techniques is by no means new, little in the way of systematic investigation of their nature or relation to other statistical techniques emerging from the Fisherian tradition has been undertaken.

In such nationally important research undertakings as the NAEP studies of educational progress, it was appropriate to employ such techniques, still without a detailed understanding of their nature. Chief among these techniques was that known as balancing, defined for situations in which the basic data are proportions of successes in the cells of a multiply classified table, usually with unequal numbers of basic observations in the several cells. It was, at the outset, known that simple comparisons of raw proportions would lead to confounded results and hence it was appropriate to employ a technique which could potentially untangle the various influences which exhibited themselves in the data.

Balancing is not, however, the only technique that has been proposed to accomplish this end. Techniques such as direct and indirect standardization, "smear-and-sweep," and the analysis of covariance have all been employed at various times for similar purposes.

It was the aim of the research herein reported to develop a better understanding of the nature and similarities of these techniques, thereby to make possible a greater appreciation of their implications for applied research. As it has turned out, there is, for many of these techniques, a single unifying approach, that of the nonorthogonal analysis of variance. By viewing them in terms of this type of analysis, unexpected insights into their nature were found. In order to accomplish this, however, more needed to be understood about the nonorthogonal analysis of variance and hance a substantial portion of the activities of this investigation was spent on a detailed study of this technique.

It was found that the nature of the nonorthogonal analysis of variance can be understood by viewing it as a comparison of competing models with the role of significance testing being simply the means for selecting the best of the competing models. It was found that both ignoring and eliminating tests were jointly necessary to accomplish these ends and that it is not always possible to select a single best model (i.e., there is the possibility, albeit rare, for an ambiguous result). Of importance for the later insights into adjustment techniques were the results on estimation which follow the selection of the "best" model, particularly those results which bear upon marginal means and the concept of weighting.

With the results from the study of the nonorthogonal analysis of variance firmly established, we looked more closely at the several adjustment techniques. As had been speculated, it was possible to show that if one defines success or failure as a binary random variable, the equations of the balancing method are identical to those that define the nonorthogonal analysis of variance in a main effects model. The result of this equivalence is that one can, with some care, use standard ANOVA programs to perform balancing; consequently, the large body of literature concerning analysis of variance can be directly applied to the balancing situation. Of greater importance for interpretation, however, is the virtual identity between the balanced estimates and the estimated marginal means in nonorthogonal ANOVA. This identity led us to explore the various types of weighting schemes for marginal means and to conclude that, in a more general, and possibly interactive context, one needs first to adopt a linear model which accurately reflects the population from which the data were obtained. Following the selection of the appropriate model (a significance testing problem) and the proper estimation of parameters in that model (an estimation problem independent of the significance testing problem), the weights are then chosen as a function of the use to which the marginal means are to be put. Balancing, which inherently implies a main effects model, has been used to compare groups as if the groups were comparable on other variables. This necessarily implies estimation in a main effects model followed by weighting with singly subscripted weights. If, however, one were to decide that an interactive model was more appropriate (by use of the nonorthogonal ANOVA, for instance), one could estimate cell means in that model and then again use singly subscripted weights to draw the same type of conclusions but under a rather different model of nature.

Direct standardization can be viewed in a similar way. Since direct standardization is based on observed cell means (estimates from an interactive model) the results of direct standardization must differ from those of balancing (a main effects model) when interactions are present. Standardized estimates can also be obtained by estimation in an interactive model combined with the use of singly subscripted weights based upon the proportion of cases in the "standard" populations. Indirect standardization, however, does not fit this type of model, and may give different results from balancing, even when no interaction is present.

Adjustment by analysis of covariance is similar to both balancing and direct standardization although on its face it appears to provide a different type of adjustment. If one considers a multifactor main effects ANOVA design, where for a particular factor one includes only the linear component in the model, the estimated cell means are identical to what would be obtained, had a covariate been used in place of the factor. In this special and somewhat limited case the balanced estimates will be identical to estimates adjusted for a covariate. One could as well generalize this result to interactive models, so that we have a class of adjustment procedures which are essentially equivalent, differing primarily in the choice of the appropriate linear model.

The choice between balancing, direct standardization, and analysis of covariance is necessarily dependent only upon which provides the most appropriate linear model. In fact, none of them may provide a parsimonious model, and we think it preferable to think of choosing the correct model in the more general context of nonorthogonal ANOVA with these special cases providing frequently chosen options. Indirect standardization would seem to be a less preferable choice.

The smear-and-sweep procedure differs markedly from the above procedures in that it is comparitively ill-defined and arbitrary; there is no well-justified rule for deciding the order in which classification variables are to be selected and how cells should be pooled. Our investigation suggests that the number of categories may substantially affect the estimated effects while the order of variables has a considerably smaller effect. In view of the arbitrariness involved, we can see little justification for the use of the smear-and-sweep procedure to meet the purposes that also may be served by balancing or analysis

In summary, a number of the adjustment techniques employed for the purpose of adjusting for initial differences among observed groups are closely related through the more general nonorthogonal analysis of variance. In general these techniques are actually a combination of three rather distinct processes: the determination of an appropriate linear model, the estimation of parameters, and the combining of estimates by a weighting scheme. Each technique (save smear-and-sweep) employs a particular combination of these, usually prescribed before the fact. A detailed understanding of how each operates relative to these processes then allows for a better understanding of its basic nature.

Publications and Presentations Supported or Partially Supported by this Grant.

Appelbaum, M.I. Problems in nonorthogonal analysis of variance. Talk presented at the Annual Meeting of the American Psychological Association, August, 1973.

Appelbaum, M.I. Marginal means in multivariate survey analysis. Talk presented at Southern Society of Multivariate Experimental Psychologists, Atlanta: April, 1975.

Appelbaum, M.I. & Cramer, E.M. Some problems in the nonorthogonal analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

Appelbaum, M.I. & Cramer, E.M. Balancing--Analysis of variance by another name. Journal of Educational Statistics, 1976, in press.

Cramer, E.M. An overview of nonorthogonal analysis of variance. Talk presented at the Annual meeting of the American Psychological Association, August, 1973.

Cramer, E.M. A nonorthogonal analysis of variance program. Journal of the American Statistical Association, 1976, 71, 93-95.

Cramer, E.M. Analysis of variance. In International Enclyclopedia of Neurology, Psychiatry, Psychoanalysis, and Psychology. Von Nostrand Reinbold, in press.

Cramer, E.M. & Appelbaum, M.I. Tukey's test for non-additivity as a test of the linear x linear component of interaction. L.L. Thurstone Psychometric Laboratory Report #125, October, 1973.

Cramer, E.M. & Appelbaum, M.I. The nonorthogonal analysis of variance--once again. L.L. Thurstone Psychometric Laboratory Report #143, September, 1975.

Maxwell, S.E. Adjusted female and male acceptance rates to five graduate programs of UNC-CH. Unpublished master's thesis, 1974.

Maxwell, S.E. & Cramer, E.M. A note on analysis of covariance. Psychological Bulletin, 1975, 82, 187-190.

Maxwell, S.E. & Jones, L.V. Female and male admission to graduate school: An illustrative inquiry. Journal of Educational Statistics, 1976, 1, 1-37.

Peng, S.S. The essence of balancing: Adjustment of group effects. Talk presented at AERA Annual Meeting, Washington, D.C., April, 1975.

Takane, &. & Cramer, E.M. Regions of significance in multiple regression analysis. Multivariate Behavioral Research, 1975, 10, 373-384.

Material based on the work performed under this grant has been presented in colloquia at

The University of Chicago

The University of Florida, Gainsville

The University of North Carolina, Greensboro

The University of North Carolina, Charlotte

Stanford University

References

Ahmann, J. S., Larson, R., Martin, W., Searls, D., Sherman, S., Rogers, T., and Wright, D. A look at the analysis of national assessment data. In W. E. Coffman (Ed.), Frontiers of Educational Assessment and Information Systems--1973. Boston: Houghton Mifflin Co., 1973, 89-111.

Appelbaum, M. I. and Cramer, E. M. Some problems in the non-orthogonal analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

Appelbaum, M. I. and Cramer, E. M. Balancing--Analysis of variance by another name. Chapel Hill, N.C.: University of North Carolina Psychometric Laboratory Report No. 143, 1975.

Bancroft, T. A. Topics in Intermediate Statistical Methods. Vol. I. Ames, Iowa: The Iowa State University Press, 1968

Bickel, P. J., Hammel, E. A., and O'Connell, J. W Sex bias in graduate admissions: Data from Berkeley. Science, 1975, 187, 398-404.

Bock, R. D. Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill, 1975.

Box, G. E. P. and Muller, M. E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29.

Cochran, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 1968(a), 24, 295-313.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968(b), 10, 637-666.

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cramer, E. M. Revised MANOVA program. Psychometric Laboratory, University of North Carolina at Chapel Hill, 1967.

Cramer, E. M. Significance tests and tests of models in multiple regression. The American Statistician, 1972, 26, 26-30.

Cronbach, L. J. and Furby, L. How we should measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Draper, N. R. and Smith, H. Applied Regression Analysis. New York: Wiley, 1966.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-401.

Evans, S. H. and Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. Psychological Bulletin, 1968, 69, 225-2[illegible].

Finn, J. D. Multivariance: Univariate and Multivariate Analysis of Variance, Covariance and Regression. Ann Arbor: National Education Resources, Inc. 1972.

Fleiss, J. L. Statistical Methods for Rates and Proportions. New York: Wiley, 1973.

Frankel, M. M. and Beamer, J. F. Projections of Educational Statistics to 1982-83. Washington: U. S. Government Printing Office, 1973.

Gentleman, W. M., Gilbert, J. P., and Tukey, J. W. The smear-and-sweep analysis. In J. P. Bunker, W. H. Forrest, F. Mosteller, and L. D. Vandam (Eds.), The National Halothane Study. U. S. Government Printing Office, Washington, D. C., 196[illegible] [illegible].

Gilbert, J. P. and [illegible]man, P. Statistical techniques used in the National Halothane Study and in the National Assessment of Education. Unpublished manuscript. December 20, 1973.

Joe, G. W. Comment on Overall and Spiegel's "Least squares analysis of experimental data." Psychological Bulletin, 1971, 75, 364-366.

Jones, L. V. The nature of measurement. In R. L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971.

Kalton, G. Standardization: A technique to control for extraneous variables. Applied Statistics, 1968, 17, 118-136.

Lord, F. M. Large-sample covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 309-321.

Maxwell, S. Adjusted female and male acceptance rates to five graduate programs of UNC-CH. Unpublished master's thesis, University of North Carolina, 1974.

Maxwell, S. and Cramer, E. M. A note on analysis of covariance. Psychological Bulletin, 1975, 82, 187-190.

Meehl, P. E. Nuisance variables and the ex post facto design. In M. Radner and S. Winokur (Eds.), Minnesota Studies in the Philosophy of Science, Vol. IV. Minneapolis: University of Minnesota Press, 1970.

Moses, L. E. Comparison of crude and standardized anesthetic death rates. In J. P. Bunker, W. H. Forrest, F. Mosteller, L. D. Vandam (Eds.), The National Halothane Study: A Study of the Possible Association Between Halothane Anesthesia and Postoperative Hepatic Necrosis. Washington, D. C.: U. S. Government Printing Office, 1969.

National Assessment of Educational Progress, Report 7: Science Group Results B, 1971. Washington, D. C.: U. S. Government Printing Office, 1973.

Overall, J. E. and Spiegel, D. K. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Overall, J. E., Spiegel, D. K., and Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. Psychological Bulletin, 1975, 82, 182-186.

Rawlings, Jr., R. R. None on non-orthogonal analysis of variance. Psychological Bulletin, 1972, 77, 373-374.

Snedecor, G. W. and Cochran, W. G. Statistical Methods (6th edition). Ames, Iowa: Iowa State University Press, 1967.

Solmon, L. C. Women in doctoral education: Clues and puzzles regarding institutional discrimination. Research in Higher Education, 1973, 1, 299-332.

Tatsuoka, M. M. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: Wiley, 1971.

Walls, R. C. and Weeks, D. L. A note on the variance of a predicted response in regression. The American Statistician, 1969, 23, 24-26.

Werts, C. E. and Linn, R. L. Causal assumptions in various procedures for the least squares analysis of categorical data. Psychological Bulletin, 1971, 75, 430-431.

Wiley, D. E. Auf dem Wege zum "Ceteris Paribus." Datenkorrektur in der Bildungsforschung. In W. Edelstein and D. Hopf (Eds.), Bedingungen des Bildungsprozesses: Psychologische und Pädagoische Forschung zum Lehren und Lernen in der Schule. Stuttgart: Klett, 1973. (Also appears as: Approximations to ceteris paribus: Data adjustment in educational research. In W. H. Sewell, R. M. Hauser, and D. L. Featherman (Eds.), Schooling and Achievement in American Society. New York: Academic Press, in press.

Williams, J. D. Two way fixed effects analysis of variance with disproportionate cell frequencies. Multivariate Behavioral Research, 1972, 7, 67-83.

Willingham, W. W. Predicting success in graduate education. Science, 1974, 183, 273-278.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.

Appendix A

DATA SET I

1	1	1	1	2,412100
1	1	1	1	2,739699
1	1	1	1	4,130899
1	1	1	1	2,334700
1	1	1	1	1,858999
1	1	1	2	1,138700
1	1	1	2	1,286500
1	1	1	2	0,733300
1	1	1	2	2,833099
1	1	2	1	0,818400
1	1	2	1	-3,617700
1	1	2	1	-2,994800
1	1	2	2	-0,181600
1	1	2	2	-1,617700
1	2	1	1	2,005200
1	2	1	1	4,315499
1	2	1	2	3,315499
1	2	1	2	4,138700
1	2	1	2	4,286500
1	2	2	1	1,467601
1	2	2	1	4,312900
1	2	2	1	2,307400
1	2	2	1	2,657399
1	2	2	2	1,307400
1	2	2	2	1,657399
1	2	2	2	4,181299
1	2	2	2	2,739900
1	2	2	2	2,104500
2	1	1	1	-2,266700
2	1	1	1	-0,166901
2	1	1	1	-3,532399
2	1	1	2	-2,861300
2	1	1	2	-2,713500
2	1	2	1	-1,818701
2	1	2	1	-3,260099
2	1	2	1	-3,895499
2	1	2	1	-3,587899
2	1	2	2	-4,713500
2	1	2	2	-5,266700
2	1	2	2	-3,166901
2	1	2	2	-6,532399
2	1	2	2	-3,687100
2	2	1	1	1,818399
2	2	1	1	0,382300
2	2	1	1	-1,994800
2	2	1	1	0,315500
2	2	1	2	1,312900
2	2	1	2	-0,692600
2	2	1	2	-0,342600
2	2	2	1	-0,181600
2	2	2	1	-1,617700
2	2	2	1	-3,994800
2	2	2	1	-1,684500
2	2	2	1	-0,861300
2	2	2	2	-2,266700
2	2	2	2	-0,166901

ATA SET II

1	1		1	1,512099
1	1		1	1,839699
1	1	1	1	3,230899
1	1	1	1	1,434699
1	1	1	1	0,958999
1	1	1	2	1,038699
1	1	1	2	1,186500
1	1	1	2	0,633300
1	1	1	2	2,733099
1	1	2	1	0,918400
1	1	2	1	-3,517700
1	1	2	1	-2,894799
1	1	2	2	0,718400
1	1	2	2	-0,717700
1	2	1	1	0,105200
1	2	1	1	2,415499
1	2	1	2	2,215499
1	2	1	2	3,038699
1	2	1	2	3,186500
1	2	2	1	0,567600
1	2	2	1	3,412899
1	2	2	1	1,407399
1	2	2	1	1,757399
1	2	2	2	1,207399
1	2	2	2	1,557399
1	2	2	2	4,081299
1	2	2	2	2,639899
1	2	2	2	2,004499
2	1	1	1	-2,166699
2	1	1	1	-0,066900
2	1	1	1	-3,432399
2	1	1	2	-1,961299
2	1	1	2	-1,813499
2	1	2	1	-0,718700
2	1	2	1	-2,160099
2	1	2	1	-2,795499
2	1	2	1	-2,487899
2	1	2	2	-2,813499
2	1	2	2	-3,366699
2	1	2	2	-1,266900
2	1	2	2	-4,632399
2	1	2	2	-1,787100
2	2	1	1	0,918400
2	2	1	1	-0,517700
2	2	1	1	-2,894799
2	2	1	1	-0,584500
2	2	1	2	1,212899
2	2	1	2	-0,792600
2	2	1	2	-0,442600
2	2	2	1	-0,081600
2	2	2	1	-1,517699
2	2	2	1	-3,894799
2	2	2	1	-1,584499
2	2	2	1	-0,761300
2	2	2	2	-1,366699
2	2	2	2	0,733100

DATA SET III

1	1	1		1,062099
1	1	1		1,389699
1	1	1		2,780899
1	1	1	1	0,984700
1	1	1		0,509000
1	1			0,688700
1	1			0,836500
1	1	1		0,283300
1	1	1		2,383099
1	1	2		1,268399
1	1	2	1	-0,167700
1	1	2	1	-2,544799
1	1	2	2	1,168400
1	1	2	2	-0,267700
1	2	1	1	-1,344801
1	2	1	1	0,965499
1	2	1	2	0,865499
1	2	1	2	1,688699
1	2	1	2	1,836499
1	2	2	1	-0,082400
1	2	2	1	2,762898
1	2	2		0,757399
1	2	2		1,107398
1	2	2	2	0,657399
1	2	2	2	1,007399
1	2	2	2	3,531299
1	2	2	2	2,089899
1	2	2	2	1,454499
2	1	1	1	-1,616699
2	1	1	1	0,483100
2	1	1	1	-2,882399
2	1	1	2	-1,311298
2	1	1	2	-1,163499
2	1	2	1	0,631300
2	1	2	1	-0,810099
2	1	2	1	-1,445498
2	1	2	1	-1,137898
2	1	2	2	-1,363499
2	1	2	2	-1,916698
2	1	2	2	0,183099
2	1	2	2	-3,182398
2	1	2	2	-0,337099
2	2	1	1	0,468400
2	2	1	1	-0,967700
2	2	1	1	-3,344799
2	2	1	1	-1,034499
2	2	1	2	0,862900
2	2	1	2	-1,142599
2	2	1	2	-0,792600
2	2	2	1	0,268400
2	2	2	1	-1,167700
2	2	2	1	-3,544799
2	2	2	1	-1,234500
2	2	2	1	-0,411300
2	2	2	2	-0,916700
2	2	2	2	1,183099

Appendix A

DATA SET I

1	1	1	1	2,412100
1	1	1	1	2,739699
1	1	1	1	4,130899
1	1	1	1	2,334700
1	1	1	1	1,858999
1	1	1	2	1,138700
1	1	1	2	1,286500
1	1	1	2	0,733300
1	1	1	2	2,833099
1	1	2	1	3,818400
1	1	2	1	-2,417700
1	1	2	1	-2,994800
1	1	2	2	-1,181500
1	1	2	2	-1,617700
1	2	1	1	2,005200
1	2	1	1	4,315499
1	2	1	2	3,315499
1	2	1	2	4,138700
1	2	1	2	4,286500
1	2	2	1	1,467601
1	2	2	1	4,312900
1	2	2	1	2,307400
1	2	2	1	2,657399
1	2	2	2	1,307400
1	2	2	2	1,657399
1	2	2	2	4,181299
1	2	2	2	2,739900
1	2	2	2	2,104500
2	1	1	1	-2,266700
2	1	1	1	-0,166901
2	1	1	1	-3,532399
2	1	1	2	-2,861300
2	1	1	2	-2,713500
2	1	2	1	-1,818701
2	1	2	1	-3,260099
2	1	2	1	-3,895499
2	1	2	1	-3,587899
2	1	2	2	-4,713500
2	1	2	2	-5,266700
2	1	2	2	-3,166901
2	1	2	2	-6,532399
2	1	2	2	-3,687100
2	2	1	1	2,818399
2	2	1	1	1,382300
2	2	1	1	-2,994800
2	2	1	1	1,315500
2	2	1	2	1,312900
2	2	1	2	-0,692600
2	2	1	2	-0,342600
2	2	2	1	-0,181600
2	2	2	1	-1,517700
2	2	2	1	-3,994800
2	2	2	1	-1,684500
2	2	2	1	-0,861300
2	2	2	2	-2,266700
2	2	2	2	-0,166901

DATA SET II

1	1	1	1	1,512099
1	1	1	1	1,839699
1	1	1	1	3,230899
1	1	1	1	1,434699
1	1	1	1	0,958999
1	1	1	2	1,038699
1	1	1	2	1,186500
1	1	1	2	0,633300
1	1	1	2	2,733099
1	1	2	1	0,918400
1	1	2	1	-0,517700
1	1	2	1	-2,894799
1	1	2	2	0,718400
1	1	2	2	-0,717700
1	2	1	1	0,105200
1	2	1	1	2,415499
1	2	1	2	2,215499
1	2	1	2	3,038699
1	2	1	2	3,186500
1	2	2	1	0,567600
1	2	2	1	3,412899
1	2	2	1	1,407399
1	2	2	1	1,757399
1	2	2	2	1,207399
1	2	2	2	1,557399
1	2	2	2	4,081299
1	2	2	2	2,639899
1	2	2	2	2,004499
2	1	1	1	-2,166699
2	1	1	1	-0,066900
2	1	1	1	-3,432399
2	1	1	2	-1,961299
2	1	1	2	-1,813499
2	1	2	1	-0,718700
2	1	2	1	-2,160099
2	1	2	1	-2,795499
2	1	2	1	-2,487899
2	1	2	2	-2,813499
2	1	2	2	-3,366699
2	1	2	2	-1,266900
2	1	2	2	-4,632399
2	1	2	2	-1,787100
2	2	1	1	0,918400
2	2	1	1	-0,517700
2	2	1	1	-2,894799
2	2	1	1	-0,584500
2	2	1	2	1,212899
2	2	1	2	-0,792600
2	2	1	2	-0,442600
2	2	2	1	-0,081600
2	2	2	1	-1,517699
2	2	2	1	-3,894799
2	2	2	1	-1,584499
2	2	2	1	-0,761300
2	2	2	2	-1,366699
2	2	2	2	0,733100

DATA SET III

1	1	1	1	1,062099
1	1	1	1	1,389699
1	1	1	1	2,780899
1	1	1	1	0,984700
1	1	1	1	0,509000
1	1	1	2	0,688700
1	1	1	2	0,836500
1	1	1	2	0,283300
1	1	1	2	2,383099
1	1	2	1	1,268399
1	1	2	1	-0,167700
1	1	2	1	-2,544799
1	1	2	2	1,168400
1	1	2	2	-0,267700
1	2	1	1	-1,344801
1	2	1	1	0,965499
1	2	1	2	0,865499
1	2	1	2	1,688699
1	2	1	2	1,836499
1	2	2	1	-0,082400
1	2	2	1	2,762898
1	2	2	1	0,757399
1	2	2	1	1,107398
1	2	2	2	0,657399
1	2	2	2	1,007399
1	2	2	2	3,531299
1	2	2	2	2,089899
1	2	2	2	1,454499
2	1	1	1	-1,616699
2	1	1	1	0,483100
2	1	1	1	-2,882399
2	1	1	2	-1,311298
2	1	1	2	-1,163499
2	1	2	1	0,631300
2	1	2	1	-0,810099
2	1	2	1	-1,445498
2	1	2	1	-1,137898
2	1	2	2	-1,363499
2	1	2	2	-1,916698
2	1	2	2	0,183101
2	1	2	2	-3,182398
2	1	2	2	-0,337099
2	2	1	1	0,468400
2	2	1	1	-0,967700
2	2	1	1	-3,344799
2	2	1	1	-1,034499
2	2	1	2	0,862900
2	2	1	2	-1,142599
2	2	1	2	-0,792600
2	2	2	1	0,268400
2	2	2	1	-1,167700
2	2	2	1	-3,544799
2	2	2	1	-1,234500
2	2	2	1	-0,411300
2	2	2	2	-0,916700
2	2	2	2	1,183099